This book is the property
of
— VAN HORN —
The Mighty Educator
— SUMMER 1928 —
424 So MAIN ST. ←
Girls - Have you never been there?
Your Loss

Library
OHIO NORTHERN UNIVERSITY

Regulations

1. Books may be borrowed for a period of two weeks.
2. At the expiration of two weeks books may be re-let if returned to the librarian in good condition.
3. A rental of two cents a day will be charged on each book held over two weeks.

Class No. 371.24 Accession No. 25107

eterick Memorial Library
hio Northern University
da, Ohio 45810

1930

G. H. Deeming

AMERICAN EDUCATION SERIES

GEORGE DRAYTON STRAYER, GENERAL EDITOR

AMERICAN EDUCATION SERIES
GEORGE DRAYTON STRAYER, GENERAL EDITOR

MEASURING RESULTS IN EDUCATION

BY

MARION REX TRABUE

DIRECTOR OF THE BUREAU OF EDUCATIONAL RESEARCH
UNIVERSITY OF NORTH CAROLINA

AMERICAN BOOK COMPANY
NEW YORK CINCINNATI CHICAGO
BOSTON ATLANTA

COPYRIGHT, 1924, BY
AMERICAN BOOK COMPANY

MEASURING RESULTS IN EDUCATION — TRABUE

E. P. 1

MADE IN U.S.A.

T75

EDITOR'S INTRODUCTION

THE only final test of the validity of methods of teaching is found in the changes which are brought about in the knowledge, habits, thinking, and attitudes of pupils. In like manner, the particular forms of school organization, and the administrative procedures which are followed, must be judged by the results which are secured. The development of the technique of measurement in these fields constitutes one of the most important contributions in education during the past few years.

Teachers have always sought to measure the results of their work. In the regular class periods and at stated times tests and examinations have been given in order to determine the progress which pupils are making. The traditional examination, at its best, was a very imperfect instrument. With the development of the technique of measurement in the field of education, it has been possible not only to determine more accurately the achievements of pupils, but also to diagnose their particular difficulties. It seems reasonable to expect that every careful professional worker will in some degree master this important instrument of his profession.

In the earlier chapters of this book Dr. Trabue has given many illustrations of the use of the techniques commonly employed in the measuring of results. In the later chapters he has established the validity of the processes of meas-

25107

urement through a discussion of their derivation, and by further illustration of their application.

The more important problems with which teachers and supervisors are constantly struggling are used for purposes of illustration. The author has in this manner gained the double advantage of making the book relatively easy to understand, and at the same time placing the student who masters it in possession of the techniques which he will employ in his everyday work.

GEORGE D. STRAYER

PREFACE

If a recognized science of instruction is to be developed, teachers and their supervisors must have a clear understanding of the purposes and methods of measuring results in education. While there are numerous books presenting selected lists of scales and tests, and a few volumes giving discussions of the theory of educational measurements, the average teacher finds that the former do not furnish an adequate interpretation of the subject and that the latter are for the most part beyond her powers of understanding. This volume is an effort to interpret the general principles of measurement for the average elementary school teacher. A few specific tests have been presented as illustrations of certain types of measurement, but emphasis has been directed in all cases to the principles involved rather than to the tests used as examples.

The outline and parts of certain chapters were written and tried out at Teachers College, Columbia University, more than eight years ago. Changes in the order and treatment of topics have been made from time to time as a result of further trials with introductory classes in educational measurements. The present arrangement has proved most useful in the writer's classes and is recommended for general introductory courses.

It is a pleasure to acknowledge the coöperation of the writer's students and colleagues at Columbia University

and at the University of North Carolina. Large contributions have also been made directly and indirectly by the research workers in all parts of the educational world. The completion and publication of the manuscript are due chiefly to the long-continued stimulation and encouragement of Dr. George D. Strayer, the general editor of the American Education Series.

M. R. TRABUE

CONTENTS

CHAPTER PAGE

I. THE NATURE OF MEASUREMENTS . . . 17

The results of education are measured by evidences that pupils are learning. Measurements are more exact descriptions and records of observed conditions.

II. SCHOOL MARKS AND PUPILS' REPORT CARDS . 34

Teachers do not agree in their purposes, methods, or abilities for assigning scholarship marks. School marks do not measure effectively any of the desired results in education.

III. HOW STANDARD TESTS DEVELOPED . . . 59

Administrative officers, attempting to defend their modern practices, gave the same tests under different conditions. Psychologists borrowed methods of statistical treatment from older sciences and made them available for educational tests.

IV. A TEST FOR THE AGES OF SCHOOL PUPILS . 80

Accuracy is necessary in the collection of data, even in such a universal measurement as that of ages. One must be conscious of the meanings of all scores employed.

V. DISTRIBUTIONS OF AGE SCORES . . . 96

Elementary statistical processes are introduced in this chapter: seriation, tabulation, and the calculation of percentile points.

VI. A TEST FOR SPELLING ABILITY . . . 128

An illustration of a measurement of results in education, showing test directions, measuring scales, and comparative scores by ages and by grades.

CHAPTER PAGE

VII. SPELLING SCORES OF A SCHOOL SYSTEM . . 144

An illustration of the statistical treatment to which scores in an educational test may be subjected.

VIII. A TEST FOR HANDWRITING QUALITY . . 172

A review of statistical operations with a different type of scale, involving different limits to scores.

IX. AVERAGE MEASURES 197

Generalized discussion of how to calculate mean, median, and modal measures under various conditions.

X. DISTRIBUTIONS OF FREQUENCY . . . 231

The causes and characteristics of normal surfaces of frequency, with the effects of inadequate samplings of total situations.

XI. MEASURES OF VARIABILITY AND DISPERSION . 252

Methods of measuring and reporting amounts of dispersion and skewness in distributions of frequency.

XII. TESTS OF SPEED IN ARITHMETIC . . . 275

Illustrated by the Courtis Standard Research Tests, Series B, and the Cleveland Survey Tests. Directions, standard scores for comparison, and suggestions for use.

XIII. TESTS OF POWER IN ARITHMETIC . . . 294

Illustrated by the Woody Arithmetic Scales, Series B, the Stone Reasoning Test, and Buckingham's Scale for Problems in Arithmetic. Complete directions for use.

XIV. CLASSIFICATIONS OF MENTAL TESTS . . . 314

Attention is called to various points of view from which measurements may be classified helpfully. The doctrine of the dominant variable is emphasized.

XV. MEASURING ABILITY TO READ . . . 339

An analysis of some available reading tests from the point of view of the abilities measured by each. An

CHAPTER PAGE

attempt to show how to study new tests as they are published.

XVI. Scales for Quality in English Composition 362

A comparison of the most important scales, with directions for their use and interpretation. Special statistical difficulties noted and simplified.

XVII. Measurements of Relationship . . . 389

Construction of correlation tables and the calculation of coefficients of correlation from them.

XVIII. Tests of Intellectual Ability . . . 412

The development of measurements for intelligence, and principles for the selection of tests to be used.

XIX. Educational Uses of Intelligence Tests . 429

Use of tests in the classification, guidance, and instruction of pupils. The importance of studying the relationships between scores in different tests.

XX. The Derivation of Educational Scales . 449

The fundamental importance of a permanent basis for determining quality and difficulty. Dispersion in a normal surface of frequency the best basis yet discovered.

XXI. Prospects in Educational Measurements . 468

Probable changes in the character of measurements as teachers take them over from the administrators. The effects of measurements on examinations, scholarship marks, methods, supervision, courses of study, and the like.

LIST OF TABLES

TABLE		PAGE
I.	Percentages of high school failures by classes	45
II.	Distributions of high school English marks by different teachers	46
III.	Distributions of mathematics marks by different teachers	48
IV.	Additional distributions of marks in English	49
V.	An age computation table	86
VI.	A cumulative frequency table	99
VII.	Eighth-grade pupils' ages	100
VIII.	An age frequency table	101
IX.	Distribution of ages of fourth-year high school pupils	102
X.	Tally method of distributing ages of sixth-grade pupils	108
XI.	Graphic tabulation of ages of sixth-grade pupils	108
XII.	A cumulative frequency table	112
XIII.	Ages of fourth-year high school pupils	121
XIV.	Grade scores on standard spelling list	138
XV.	Age scores on standard spelling list	138
XVI.	Distribution by grades of spelling scores in Baltimore	149
XVII.	Calculation of the mean point in a distribution of scores	153
XVIII.	Averages and dispersion of spelling scores in elementary grades	157
XIX.	Distribution by ages of spelling scores in Grades IV to VIII	159
XX.	Dispersion of spelling scores in Baltimore by ages	160
XXI.	Decile points in spelling scores for elementary grades	166
XXII.	Medians of spelling scores by ages within each grade	167
XXIII.	Quality of handwriting in four cities by grades	181
XXIV.	Quality of handwriting in two cities by ages	182
XXV.	Median quality of handwriting under varying conditions	184
XXVI.	Handwriting objectives for the elementary grades	186
XXVII.	Handwriting scores of twelve-year-old sixth-grade pupils	188
XXVIII.	Calculation of a mean handwriting value	192
XXIX.	Daily record of handwriting speed test results	199
XXX.	Frequency distribution of handwriting speed scores	200

TABLE		PAGE
XXXI.	Spelling scores of fifteen-year-old pupils	205
XXXII.	Fifth-grade handwriting scores	206
XXXIII.	Calculation of mean speed score in handwriting	211
XXXIV.	Calculation of mean fifth-grade handwriting quality	211
XXXV.	Calculation of mean spelling score of fifteen-year-old pupils	213
XXXVI.	Calculation of mean speed score in handwriting	215
XXXVII.	Distributions of sixth-grade scores	217
XXXVIII.	A distribution of scores for quality of handwriting	221
XXXIX.	Last six speed scores in handwriting test	226
XL.	Calculation of fifteen-year-old spelling mean	227
XLI.	Calculation of mean deviation from tabulated data	261
XLII.	Handwriting scores of fourteen-year-old pupils in eighth grade	263
XLIII.	Calculation of standard deviation from tabulated data	266
XLIV.	Calculation of standard deviation from an assumed mean	268
XLV.	Median number of problems solved correctly in Courtis Tests	286
XLVI.	Average scores by grade in Cleveland Survey Arithmetic Tests	290
XLVII.	Difficulty value of each problem in Woody Arithmetic Scales, Series B	299
XLVIII.	Comparative scores by grades, Woody Arithmetic Scales, Series B	302
XLIX.	Median scores by grades in Stone Reasoning Tests	307
L.	Mid-year standards in Buckingham Scale for Problems in Arithmetic	311
LI.	English composition results by age in each elementary grade	377
LII.	English composition results by age and by grade	378
LIII.	Standard English composition scores by grades	379
LIV.	Calculation of percentile values on Hillegas Scale	383
LV.	Calculation of percentile values on Nassau County Supplement	383
LVI.	Correlation examples: Scores of nine pupils in six tests	393
LVII.	Correlation between tests A and D	396
LVIII.	Correlation between tests A and C	397
LIX.	Correlation between tests A and E	399

TABLE		PAGE
LX.	Correlation between intelligence and reading	401
LXI.	Correlation between scholarship and conduct marks	404
LXII.	P. E. of difference between true and obtained correlations	405
LXIII.	P. E. distances from the median corresponding to per cents of the normal surface of frequency	457

LIST OF FIGURES

FIGURE		PAGE
1.	Household thermometer. A scale for measuring temperature	19
2.	Blank for recording a pupil's scholarship marks	37
3.	Differences between high schools in assigning a grade of Failure	44
4.	Differences between English teachers in the assignment of grades	47
5.	Differences in the marking scales used by different teachers	51
6.	Marks assigned by 91 teachers to English paper A	54
7.	Marks assigned by 91 teachers to English paper B	54
8.	Marks assigned by 70 teachers to the same history paper	55
9.	Four ways of grouping ages	90
10.	Graphic comparison of distributions of seventh- and eighth-grade ages	103
11.	Graphic comparison of fourth-year high school ages in two cities	104
12.	Surface of frequency representing scores of individual pupils	106
13.	Graphic representation of sixth-grade ages	109
14.	Percentile values of year limits on measuring scale for age	114
15.	Values of the decile and quartile points in a distribution of age scores	118
16.	Comparison by deciles of two distributions of age scores	119
17.	Decile curves for two distributions of age scores	119
18.	Assumption that all scores are distributed evenly over a step	122
19.	Unequal difficulty of words given same weight in spelling test	130
20.	Comparative scores in spelling, by school grades	139
21.	Comparative scores in spelling, by ages	139
22.	The range of a spelling score of 10	146
23.	Range of the middle half of the spelling scores by grades	158
24.	Surfaces of frequency for Baltimore spelling scores, Grades IV–VIII	162
25.	Decile curves for Baltimore spelling scores, Grades IV to VIII	165
26.	Lower abilities of older pupils in each grade	168
27.	Selected samples from the Thorndike Handwriting Scale	179
28.	Quality of handwriting rated to nearest sample on scale	180
29.	A fifth-grade boy's speed scores in handwriting	201

FIGURE		PAGE
30.	Surface of frequency showing heights of adult Englishmen	234
31.	Normal surfaces of frequency	235
32.	Theoretical results of tossing coins	238
33.	Actual results of tossing coins	239
34.	Effect of small numbers of cases on form of distributions	244
35.	A distribution having positive skewness	245
36.	A negatively skewed distribution	246
37.	A frequency polygon, skewed to the low end	247
38.	Smoothing irregular curves mathematically	249
39.	Differences in dispersion of two groups of pupils	253
40.	Distributions differing in dispersion and in average	256
41.	The meaning of a plus 1 deviation	258
42.	Class record sheet for Courtis Research Arithmetic Tests, Series B	279
43.	Correlation tables showing various degrees of relationship	394
44.	Schematic representation of a variety of intellectual interests	433
45.	A graphic chart for showing educational test scores	433
46.	Lack of uniformity in differences between decile points	544
47.	Difficulty locations of two addition problems in Grades II and III	459
48.	Overlapping of one distribution of scores on another	460
49.	Overlapping of distributions of judgments on two English compositions	464
50.	Illustrative distributions of scholarship marks	484

MEASURING RESULTS IN EDUCATION

CHAPTER I

THE NATURE OF MEASUREMENTS

EVERY detail of educational organization and each item of expense in connection with the public schools must find its final justification in the contribution which it makes, either directly or indirectly, to more effective learning on the part of the pupils. A school may be housed in a well-constructed building and have all the latest maps, charts, desks, apparatus, and equipment; it may have well-trained teachers, supervisors, and administrative officers; its courses of study may be thoroughly modern and fully approved of by the best educational authorities; but unless its pupils are learning effectively it may not rightfully be called a "successful" school.

Teachers are inclined to think of instruction as the most important activity of the schoolroom. It is certainly the most important activity of the teacher. There can be no real teaching, however, unless some one is learning. It is no more absurd to speak of a farmer selling corn when no one is buying the corn than it is to speak of teaching frac-

tions when no one is learning fractions. Teaching and learning are so closely related that it is impossible for the former to exist without the latter.

The efficiency of instruction is in direct proportion to the success of the learning that is taking place in the pupils. In general, it may be said that pupils learn most rapidly and effectively when taught by trained teachers who have had some experience. But occasionally a girl just graduated from the normal training school, without any practical experience, is more efficient, if measured by the success of her pupils as learners, than other teachers who have taken Master of Arts degrees and have had many years of practical experience. The amount of training a teacher has had, the approved methods of instruction she tries to use, and the popularity she has among her students cannot be taken as conclusive evidences of successful teaching. The only convincing proofs of successful school work must be found in the increasing achievements of pupils when meeting recognized tests of ability.

Evidences of learning. — Learning is essentially a matter of *changes* in the abilities of the learner. The pupil may increase the number of facts he can repeat from memory, he may increase the speed with which he is able to add small numbers, or he may improve his skill in sawing a board. The learning may consist in so changing his tastes that he comes to like violin music more than the noise of a bass drum. But in any case of learning the essential element is a change from one state or condition to another.

To prove that learning has taken place, it is necessary to compare the learner's ability before he was taught with his ability after being taught. If no change has taken place,

then no appreciable amount of learning has occurred. The amount of difference between the ability before and the ability after being taught is the most reliable index of the amount of learning that has occurred. The amount of learning is proportional to the amount of change in ability.

Evidences of ability. — The best evidences of ability are the results one actually accomplishes in any line of activity when giving his whole thought and attention to it. No one is able to look directly into a pupil's nervous and muscular system to determine from the excellence of the organization therein what the pupil's ability in any particular line may be. One can only assume that, when the pupil is doing his best at a task, his results are a fair index of his ability.

Fig. 1.—Household thermometer. A scale for measuring temperature.

This measurement of mental abilities by means of their results is not far different from many of the common measurements of physical facts. When one wishes to know the temperature of the air or of a liquid, one measures the height of a column of mercury in a thermometer. The height of the column of mercury is a direct result of the temperature, in much the same way that what one actually accomplishes is the result of one's ability. Similarly, one measures time by the location of a shadow on a sundial or of the hands on a clockface, and one measures weight by the location of a movable piece of metal on a carefully balanced and graduated scale.

Ability is a variable quantity. — The measurement of any human ability presents difficulties much greater than

those usually encountered in measuring the weight of a stone or the length of a board, because human qualities are much less constant than the characteristics of a stone or of a board. A board which is ten feet long to-day will be ten feet long to-morrow or next month, but a boy who solves ten problems correctly in two minutes to-day may not be able to solve more than nine problems of the same difficulty in two minutes to-morrow. Mental abilities are influenced by physical growth, rest, sleep, illness, and many other factors which do not influence the characteristics of inanimate objects. The strength of a human ability varies much more than the amount of weight, length, color, or strength of a physical object.

Variability in physical measurements. — Even such lifeless objects as those mentioned above do not remain absolutely constant in all their characteristics. The steel rails used on a railroad expand during warm weather and contract when it is colder. This may be seen by observing in the winter that there is sometimes as much as a half inch of space between adjoining rails which in the summer were in direct contact with each other. Builders of large buildings and bridges always allow for this variability in the lengths of their materials. It is probable that if one were able to make approximately perfect measurements one would find that no characteristic of any known object remains absolutely constant.

Not only do actual changes occur in objects, but two persons measuring the same object may disagree about its exact size, even though their measurements are made under almost exactly the same conditions. Different persons may interpret the scale in somewhat different ways

and therefore obtain different results. Persons who agree perfectly as to the meanings of the units on the scale may apply the scale by unlike methods and thereby obtain results which do not agree perfectly. In measuring length there may be some lack of agreement as to the extreme limits of the object being measured. In view of such variations in the measurements of the characteristics of inanimate objects, one need not be surprised at finding exactly the same types of disagreements in measurements of mental traits.

The standard yardstick. — Measurements of mental characteristics are passing through almost the same stages of development as measurements of length and weight went through centuries ago. The first units of measurement in any field are commonly very crude and inexact. In English long measure, for example, the "palm" or the "hand" was for many decades a common unit of measurement, and even yet it is customary in some places to report the height of a horse in such terms as "eighteen hands high."

Practically all peoples have used the "foot" as a unit for measuring length or distance.[1] Such natural measuring instruments, however, were not satisfactory for the reason that not all feet were the same length. Measurements made by a man with big feet were not comparable to measurements made by a man with small feet. In cases of disagreement, therefore, the king's foot came to be used as a fair compromise, upon the length of which both parties

[1] The ancient Roman "foot" was approximately 11.65 inches, while the ancient German "foot" was about 13.11 inches in length. The modern Polish "foot" is only 11.325 English inches in length, while the modern Bohemian "foot" measures 12.45 inches.

to a dispute might agree. Provision by the government of " standard weights and measures " is the modern equivalent of the king's foot.

The standard yardstick provided by the present English government as a unit of long measure is said to have been the length of the arm of King Henry I. Whether this is true or not does not matter now. The important fact is that any merchant or other person who uses a yardstick that is not the same length as the standard yardstick is using a false measure. All yardsticks must conform in their lengths to the length of the standard which is preserved by the government. The preservation of this standard is the only method by which we can be sure that one yard at the present time is exactly the same length as one yard was a hundred years ago, or the same as a yard will be a hundred years in the future. Whenever any doubt arises as to the validity of any yardstick, one may verify or correct it by comparison with the government standard.

The standard yardstick is preserved by the government scientists with great care, so that no accident may bend it or change its length. The methods and conditions for reading its length are most carefully prescribed, so that a competent physicist in any future age may use it exactly as those in the present and in the past have used it.[1] In

[1] The English standard yardstick is a bar of Bailey's Metal, 38 inches long and one inch in each of its other two dimensions. This metal was used because it was less subject to changes in temperature and to the corrosive action of air and moisture than the other metals known at the time the standard was prepared. (The appearance of Bailey's Metal is very much like what we now call gun metal.)

At one inch from each end of the bar, a hole was drilled somewhat more than half through the bar, and into each hole a gold plug was

spite of all the minute directions for its use, it is said that different scientists differ from each other, in their readings of its length, by at least one millionth of the distance from one end of it to the other. Every possible effort is made to define the standard and its use in such a manner that all competent persons will secure the same results, but when greatly refined measurements are attempted, there are slight variations in the results, even in this well-developed and "perfectly definite" field of measuring distances. With such variations occurring in physical measurements, one may expect still greater variations in the results obtained by the much newer and less well-developed measurements of mental characteristics in human beings.

Quantity in all qualities. — In ordinary conversation people speak of measuring instruments as being "exact" or "inaccurate," very much as though there were only these two degrees of accuracy. As a matter of fact, even the most perfect instruments for measuring the physical

inserted in such a way as to leave the face of the plug exactly at the central plane of the bar. On the face of each plug a line running perpendicular to the length of the bar was engraved to show the limit of the standard yard.

In order to equalize the effects of temperature, it is prescribed that the standard yard shall be the distance between the two engraved lines on the gold plugs, when the bar is suspended very carefully according to an elaborate set of directions; when the temperature is at 62 degrees Fahrenheit, according to a thermometer constructed and standardized by certain rules which are carefully specified; and when the locations of the fine, engraved lines are determined carefully, under microscopes of a specified type.

For a very interesting description of the care and use of the Standard International Meter, the reader is referred to E. A. Kennelly, "A Modern Mecca," *The Scientific Monthly*, XV, pp. 571–580.

qualities of metal or other constant objects are not "absolutely accurate." Some instruments are more nearly perfect than others, but in each case it can only be said that the instrument is exact to such and such a degree. The standard English yardstick is certainly more accurate than the yardsticks of hard wood that may be purchased from school-supply houses; and these in turn are usually more nearly accurate than the soft-wood yardsticks, which are given away as advertisements by many of our American hardware and furniture stores. And yet competent scientists, as already stated, differ somewhat even in their readings of the standard yardstick.

The same distinction must be made between the different degrees of any characteristic in any object. People may not, for example, be divided into two groups, old and young. There are many degrees of age. In the same way, there are all degrees of beauty, of honesty, of weight, of intelligence, of ability to spell, of ability to write English compositions, of ability to bake apple pies, or of any other quality or ability. It is understood generally that when some one speaks of a "tall man" he means to indicate that the man is of more than average height. No one would assume that it implies that all tall men are of equal height, or that there are only two kinds of men, tall ones and short ones. The world's population is sometimes spoken of as though there were just two groups, the good and the bad. As a matter of fact people are of all degrees of goodness, just as they are of all degrees of weight, or of intelligence. The amount of a quality possessed by any object is always a relative matter. If the quality exists at all, it has a certain strength or degree.

Although it is useful for scientific purposes to have a list of all the qualities possessed by any object that may be studied, it is still more useful to know the amount of each quality in the object. To know that a stone weighs forty pounds and three ounces is much more useful than to know that it is a " heavy " stone. To be told that a board is one inch thick, six inches wide, and sixteen feet long is much more useful to a carpenter than to be told that it is a " long " board. In order to be most useful, any report of a quality observed in an object should be accompanied by a careful measurement of the amount of the quality present.

" Standard " is a relative term. — The word " standard," when applied to a test or scale in the field of education, is just as much a relative term as is the word " long " when applied to a board in a carpenter's shop. Some scales are much more thoroughly standardized than others. One scale may be almost perfectly standardized, in so far as the selection of its subject matter is concerned, but be in no degree standard as regards its arrangement, its methods of scoring, its interpretation, and its uses. A scale may be standardized relatively well in some two or three respects, but be poorly standardized in a dozen other respects. At the present time there are very few scales in educational work that are well standardized in all respects.

Tests and scales. — There has been a great deal of discussion at educational meetings and elsewhere about " standard tests and scales," with little or no suggestion that tests and scales differ from each other in any essential particular. Where a distinction has been made, it has most frequently been a discrimination between the more and the less refined scales, the less refined being called

" tests," while those derived more accurately were called " scales." The distinction which seems most useful, and which will be used in this volume, reserves the word "scale" to indicate the measuring instrument itself, and uses the word " test " to indicate the procedure by which the child is caused to perform certain acts which may be measured on the scale.[1] The instrument by which the quality of a child's composition is to be rated would be called a " composition scale," while the set of directions, specifying just how the composition was to be assigned, what the topic and time limits were to be, and what kind of paper and pen the child should use, would be called the " composition test." The test is the procedure by which one secures a sample of the child's work to be measured on the scale.

In order to secure the most reliable results, both the test and the scale which measures the result of the test, should be well standardized. Unless the results measured by a scale were obtained by an equitable test, they are not valid as indices of the relative abilities of pupils. And unless the results of the test are measured on a reasonably valid scale, they cannot be used as trustworthy indications of the relative achievements of the pupils tested.

The nature of measurements. — Measurements in any field are essentially more exact definitions and descriptions of the amounts of various qualities possessed by objects in that field. Exactness in the definition and identification of an amount of a characteristic is made possible by comparing the amount in question to the amount contained

[1] This distinction has been very clearly made by Dr. May Ayres Burgess in *The Measurement of Silent Reading*, Russell Sage Foundation, p. 96.

in some well-known object. If the principal of a school, discussing the character of a certain boy, tells the teachers that this boy "is almost exactly as deceitful as Harry Jones," whose degree of deceitfulness is known to all the group, the principal may be said to have reported a rough measurement of the boy's character in respect to deceitfulness. Such a report may be called a "measurement," because it is relatively more exact and objective than the statement that the boy "is a very deceitful youth." The use of the well-known Harry Jones as a measure of the amount of deceitfulness supplies the necessary objective unit.

If the individual used as a measure of amount is sufficiently well known to all concerned, measurements reported in terms of that individual as a unit may be quite useful. William Jackson may, for example, be reported as having just a little more of deceitfulness in his nature than has Harry Jones. Such a statement is approximately as accurate and useful as the statement would be that John Smith is about six times as tall as the length of the king's foot. If the individual with whom comparisons are made is not well known or easily available for reference, comparisons or measurements in terms of that unit are, of course, practically useless.

Where all those concerned in the measurement are well acquainted with several individuals possessing the trait to be measured, but each possessing it in different amount, these well-known individuals may be used as rough steps on a scale for the trait in question. If Mary has greater ability as a student than Jane, while Jane has greater ability than Susan, and Alice has less than Susan, these four girls

may be used as a rough Scale for Ability-as-a-Student, by any group of teachers that knows them well. A new

SCALE FOR ABILITY-AS-A-STUDENT

(*Little Ability*) ....ALICE....SUSAN....JANE....MARY.... (*Great Ability*)

student might then be reported as having "ability about equal to Jane" with some assurance that the other teachers would understand that the new student had less ability than Mary but greater ability than Susan and Alice.

The distinguishing characteristics of measurement are its relative exactness and objectivity. A measurement is a description of an amount of some quality in terms of an objective unit of that quality which is relatively well known. Unless the amount of the characteristic contained in the reference unit is fairly objective, — that is, unless the amount may be observed and studied by any one interested in it — the report will not be clearly understood and will not be exact enough to deserve the name of measurement. The more objective the units in the scale, the more exactly may comparisons be made with the scale, and the more deserving will the results be of the term measurements.

The advantages of measurements. — The relative objectivity and exactness, which distinguish measurement from the ordinary description of characteristics, are probably the two greatest advantages of measurements. The growth and development of every science has followed rather than preceded the establishment of measured observations. Astronomy did not develop until methods had been devised for recording objectively the locations of the stars at definite hours on definite dates. Biology was not

recognized as a science until objective records began to be made of measured facts observed by nature lovers. Medicine became a science after definite records were kept of just how much of a given remedy was applied and exactly what happened in each case. Similarly, education will not be recognized as a science until it is possible to measure many characteristics of a pupil before and after a given type of instruction, and to determine thereby what is the effect of a particular treatment.

Measurement in terms of an objective scale makes it possible to convey to other people in a very short statement a relatively clear conception of the amount of a quality present in a given object. Without some objective scale by which to identify this amount of the quality in question, it is necessary to describe at great length the amount discovered; and when the description is finished the hearer may not have obtained a clear idea of how strong the characteristic actually is. The most gifted writers of descriptive articles are usually the ones who draw the best analogies and make the most discriminating comparisons between the object observed and the common experiences of their readers, — the writers who come nearest to measuring the qualities of their objects by scales known to their readers.

The effects of an observer's personal feelings and habits of thought tend to be greatly reduced by measurements. Many of us have had the experience of revisiting during adult life places we had not seen since early childhood, and have been shocked by the apparent shrinking up of distances and buildings that had previously seemed large. If we had measured these buildings and distances in our early

experiences and also in our later visits, we should not have found any actual changes of the sort that seemed to have occurred. The average Frenchman would be called " tall " by the Japanese, but would probably be considered somewhat " short " by the Norwegians, although his actual height would be the same regardless of who measured it. The elimination of errors due to personal experiences and standards, or at least their reduction to the smallest possible amounts, is one of the most important advantages of measurement over description.

Measurement has also a decided advantage in that it makes possible a comparison of objects which cannot be brought into immediate juxtaposition. As long as the scale remains the same, objects which are thousands of miles from each other may be compared with reasonable accuracy. Growing objects may be compared at different periods of their growth merely by consulting the records of their measurements. The abilities of school pupils at the present time may be measured and compared with measurements of the same abilities in pupils a hundred years from now. Comparison by means of recorded measurements rather than by actual proximity is an invaluable aid in education, where we are primarily concerned with the *changes* that occur in individuals.

Summary. — Tests and scales in educational work may no more truthfully be divided into two classes, " standard " or " non-standard," than pupils may be classified as either " tall " or " short." Every quality possessed by any person or object is possessed in some degree, which can only be called " great " (or " small ") according as it is compared with an amount of the same quality in other

persons or objects possessing less (or more) of the quality. "Standard classroom tests" are relatively more scientific and equitable ways of securing samples of classroom products, and of applying the relatively more objective and highly standardized scales for measuring their qualities. "Measurement" of a quality is more exact description of its amount, in terms of highly standardized scales, applied in accordance with the rules of well standardized tests.

"Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it — things made, words spoken, acts performed, and the like. To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement. To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used."[1]

QUESTIONS AND EXERCISES

1. What should be the chief educational justification for the use in a public school of a particular series of textbooks in arithmetic? What minor considerations might also be allowed? Do these minor points contribute in any way to the major consideration?

[1] Edward L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," p. 16, *Part II, Seventeenth Yearbook of the National Society for the Study of Education.*

Concerning the above statement, Dr. Thorndike, who is entitled to chief credit for the vitality of the present movement toward the improvement of classroom tests, makes the following remark: "This is the general *Credo* of those who, in the last decade, have been busy trying to extend and improve measurements of educational products."

2. Is it possible for one to learn without being taught? What of the child who learns that it is painful to touch a hot stove? Do text-books teach?

3. How would you attempt to determine whether a child was growing heavier? Whether he was growing taller? Whether he was writing more legibly? Whether he was growing in ability to spell?

4. By what evidences might one measure the relative lung capacities of two boys? By what evidences might one compare two sleepy boys to determine which was the sleepier? Why does one have to measure such conditions by their evidences rather than measure the conditions themselves?

5. Which is more variable from hour to hour in its amount, the desire to sleep or the capacity of one's lungs?

6. Do all those persons who look at thermometers on a hot day or on a cold morning report the same temperature? Why?

7. Measure the size of page used in this book, recording your measurement to the nearest 64th of an inch. Compare your record with that of others. List the causes for differences in results.

8. How would you select a thermometer in order to have the greatest possible confidence in its readings, if it were not possible to purchase one that had been thoroughly verified by comparison with the government standard?

9. How many inches in a "span"? To what other units of measurement is it comparable?

10. How can one distinguish a "good" person from a "bad" one? Can you draw a dividing line between a useful person and a useless one? Is there a point in its standardization where one may first begin to speak of a "standard scale"?

11. Should a foot rule be called a test or a scale? Is a bushel basket a test or a scale? Is a hundred-yard dash a test or a scale? Is a baseball game a test or a scale?

12. How objective and exact does a test need to be in order to have its results called "measurements"?

13. Are ordinary measurements of weight and length free from errors of judgment on the part of the measurers? What is meant by the "personal equation" in measurements?

REFERENCES FOR READING

C. A. Gregory, *Fundamentals of Educational Measurement*, Chapters I and II.

W. S. Monroe, *The Theory of Educational Measurements*, Chapter II.

E. L. Thorndike, *Individuality*.

E. L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," *Seventeenth Yearbook of the National Society for the Study of Education*, Part II, Chapter II.

CHAPTER II

SCHOOL MARKS AND PUPIL REPORT CARDS

THE most widely used form of measurement in education is the classroom examination. Practically every teacher, from the primary grades through the university, attempts, from time to time, to discover by means of examinations what progress is being made by his pupils. Many teachers base their estimates of individual pupils almost altogether upon the written examinations held in their classes. A considerable number of teachers conduct their recitations by the oral examination method. No other attempt at measurements in education is so universal as the attempt to determine by means of questions and answers the quality of work pupils are accomplishing.

Just as classroom examinations are the most widely used form of educational measurements, so scholarship marks are the most widely used form of record and report for measurements in education. Whether a particular teacher uses examinations extensively, sparingly, or not at all, he usually keeps no more definite records of each pupil's work than a series of letter grades or percentage marks. It would therefore seem desirable to give consideration, in any discussion of educational measurements, to the effectiveness of scholarship marks.

A boy's marks in writing. — A certain boy, well-known to the author, had his attention called to the extremely

poor quality of his handwriting. As soon as he realized the importance of good writing, the boy undertook to improve in this respect. It was pointed out to him that he could keep his copybooks, and that, by comparing the writing in those at the beginning of the year with the writing in those at the end, he could see for himself how much improvement he had made. By earnest endeavor the boy succeeded in getting his writing to conform more and more closely to the perfect copy supplied at the top of the page in his copybooks, which he kept from month to month as evidences to himself that he was improving his writing.

The report card sent home to his parents each month showed, however, almost no change throughout the entire year in his marks for handwriting. In spite of the child's belief, and copybook proof, that he was making progress in this subject, his reported grades remained approximately 80% from the first month to the last.

At the beginning of the next year, when a new teacher had charge of his instruction, the child's attention was directed toward arithmetic as the subject in which he might well strengthen himself. The writing grade sent home to his parents on the first month's report card was 75%. The next month his grade was 80% in writing, and on the third report he received 83% in this subject, in spite of the fact that he was making no effort to improve his writing.

Just as a matter of curiosity, his copybooks for the three months were brought together and compared. So far as any one could tell from the writing in the copybooks, there had been no improvement in writing between the first days of the new year and the last days of the third month. Throughout the year the grades in writing con-

tinued to improve, until the last month's report showed 90% in this subject, although the copy being written by the pupil was not approaching the models at the top of the page any more nearly than at the beginning of the year. The child's friends were forced to conclude that, during the two years considered, neither of the child's teachers had based her marks for writing on the quality of the actual penmanship of her pupils.

The purposes of reports to parents. — It is impossible to judge of the success or failure of a school procedure without knowing its purpose. The purpose of teachers and of school administrators in sending reports to parents is usually to secure more effective parental coöperation with the school. The normal parent wishes to know how his child is progressing, and thinks that the report card tells him something about the child's success. School officers send the report cards partly to satisfy the curiosity of parents, but chiefly in order that parents may help and encourage pupils to get their lessons well and "make a good showing on the report card next month."

The case of the child whose experiences with writing marks were described above seems to indicate that the information carried on report cards does not always mean just what the parents interpret it to mean. For this particular pupil, at least, a real improvement in handwriting skill did not result in higher marks, and during the following year higher marks were obtained without any accompanying improvement in skill. The child's parents would have been unjust if they had taken the 80% grade received month after month as an indication that the child was not improving his writing, and they would have been equally

mistaken had they assumed that the increasingly higher grades received during the second year indicated a real improvement in handwriting.

What do the grades sent to parents mean? If reports indicating higher grades from month to month cannot be taken as an indication of improvement in the child's work at school, how is a parent to interpret the marks appearing on report cards? If the grade received one month is not related to the grades previously received by the same pupil, there can be very little meaning in school grades for the average parent. The teacher in giving a grade might intend to show the relative *rank* of the pupil, as compared with other members of the class, but it is quite certain that no such meaning is conveyed to the parent by the grade on an ordinary report card. Such a card carries no information regarding the grades received by other pupils. If the teacher wishes,

RICHMOND PUBLIC SCHOOLS
PUPIL'S REPORT CARD

School, Grade

Pupil

Term Ending

Teacher.

RATINGS: A-90% or above. B-75% to 90%. C-below 75%. D-Indicates Deficiencies.

	MONTHS				
	1st	2nd	3rd	4th	Av.
SCHOLARSHIP					
CONDUCT					
Days Present					
Days Absent					
Times Tardy					
STUDIES					
Spelling					
Reading					
Language					
Grammar					
Arithmetic					
Geography					
History					
Hygiene					
Civics					
Writing					
Manual Training					
Music					
Nat. Study					
Drawing					
Sewing					
Cooking					
Shop Work					

FIG. 2. — Blank for recording a pupil's scholarship marks.

therefore, to have the parent urge the child to greater efforts when he is doing poorly and praise the child when he is doing well, she will report each month grades that show the relation of the child's work for that month to his work in previous months.

The purposes of scholarship records. — Assuming that the reports to parents are intended to indicate primarily the relative success of a pupil in one month as compared to his own success in previous and later months, should these grades be entered in the school records? At the present time few schools keep any record of a pupil's scholarship, other than a transcript of the grades reported to the parents. Whether such a record of scholarship is adequate depends upon its efficiency in serving the purposes to which scholarship records may legitimately be put.

The most common purpose for scholarship records within a school is to serve as a basis for the classification of pupils into groups and recitation sections. Promotion from one grade to another in most schools is dependent upon having a " passing grade " in certain studies. If scholarship grades are to be determined with reference to a pupil's own previous marks, rather than with reference to the abilities of other members of his class, these grades cannot be used effectively as a basis for making distinctions between pupils. On the other hand, it seems almost impossible to make parents understand that a grade on a report card in February has little relation to the grades for January or December, but is each month only a statement of the pupil's standing in relation to other members of the group in which he studies.

If a system of grading could be devised, whereby a single

mark would indicate both the pupil's standing with regard to his own previous performances, and at the same time his standing with reference to the other members of his class, such a combined mark might be effectively used both as a report to parents and as a basis for the classification of pupils in school. The present practice of trying to make a single percentage or letter grade indicate both of these comparisons (and frequently a considerable number of other characteristics, such as effort, intelligence, good behavior, and special aptitude in the subject) results in such a difference in the meanings of the school marks assigned by different teachers that these marks are not of great usefulness either as reports to parents or as school records of success.

A second use for scholarship records is the educational and vocational guidance of pupils. Practically every college requires those who apply for admission to submit a transcript of the marks received in their secondary school work. School and college authorities who have the responsibility of advising students turn almost universally to the scholarship records of each individual for such evidences as can be found there of general and special aptitudes or disabilities.

A third use for scholarship marks is as evidence of character and effort on the part of pupils. Administrative and supervisory officers, having problems of discipline to consider, turn frequently to the scholarship records for indications that the offenders are not maintaining the academic standards of which they are capable.

A fourth use to which records of scholarship might be put is the comparison of the relative effectiveness of differ-

ent teaching methods or of different teachers. If records of the pupils taught cannot be used to determine which of two methods of teaching is better, or which of two teachers using the same method obtains the higher results, there is little hope for a science of teaching. Unless it is possible to prove that one attempt at teaching is more effective than another, we shall never lift educational discussions above the realm of debate and opinions. Without proof on either side of an argument, it would be impossible for any one to choose the truth with assurance that it was not falsehood.

Scholarship records based on a pupil's own previous marks and achievements would not be satisfactory evidences of greater growth in one situation than in another, for the judgments of the same identical teacher may not be based on the same facts on all occasions. Then again, a pupil's mark might be made high at a given time merely because he had been " loafing " in the period just previous. Furthermore, the same amount of improvement in work may secure two or three times as much improvement in the marks given by one teacher as would be obtained from another teacher.

Scholarship records based on the comparative strength of other members of the class are somewhat more satisfactory for the school's purposes than are those based on the previous record of the same pupil. Neither basis is sufficiently free from the inaccuracies of the judgments of teachers to warrant their use in comparing different methods or different teachers. The lack of a definite and unchanging standard or point of reference makes scholarship records, based on judgments or comparisons made by teachers, en-

tirely unsatisfactory as measures of the effectiveness of a school or of a school activity.

In order to be really useful as a means of evaluating different school practices, different teaching methods, or different teachers, or in order to be truly helpful in vocational and educational guidance and in the classification of pupils according to their abilities to do school work, scholarship must be recorded in terms that are as free as possible from the opinions of teachers. Records must be kept in terms of some definite, permanent, outside standard which may be observed and used equally well by all who are interested. In other words, pupils must be measured and records must be kept in terms of objective and easily verified scales, rather than in terms of the abilities of the pupils themselves or of the abilities of their fellow-students.

The interpretation of school marks. — How much information would be supplied to the average reader by the statement that a certain pupil received a grade of B at the completion of a half year of school work?

One recognizes at first glance that certain other items of information must be supplied before any clear understanding of the meaning of the B can be obtained. The following questions must be answered before any definite impression of the situation can be formed: In what subject was the B received? To what grade or class did the pupil belong? In what type of school was the mark received? In what city or school system was the school located? How old was the pupil? Was the pupil a boy or a girl? Who was the teacher giving the mark? Was the mark given at the end of the year or at the end of the first half year? How many other pupils were in the class?

Without answers to these and many other questions it would be impossible to form any clear idea of the meaning of the B mark.

In order to compare the scholarship of Harry Jackson and Benjamin Smith, both of whom have been marked B in arithmetic, one would have to know a tremendous number of additional facts regarding the two boys, also regarding their teachers, the schools in which they worked, the course of study in arithmetic in each of the two schools, the methods by which arithmetic was taught in each school, the examination questions used by each teacher, the plan employed by each teacher in making up the marks to be assigned to her pupils, the relative intellectual abilities of the pupils in each city, and the amount of ambition possessed by each pupil for getting his work well done. It is quite evident that no report card can begin to supply this information in a single letter mark or percentage grade.

A grade of B in arithmetic cannot effectively tell the parent who examines the report card, or the principal who examines the school record, the answer to any of the following important questions: What kind of problems can the child solve correctly? How rapidly can the pupil solve problems of a given difficulty? How difficult a problem in arithmetic can the child solve if he has all the time he wants? How does the child compare with other members of his class in any of these respects? How does he compare with his own previous record? How much improvement has the pupil made during the past month or year? Is he doing his best in arithmetic? Should he be commended as having done well or should he be told that his work needs improvement?

Lack of uniformity in marking practice. — Few normal schools or colleges make any serious efforts to train prospective teachers how to determine and assign scholarship marks. Practically every teacher, trained or untrained, must work out her own scheme for assigning scholarship marks. It is not strange, therefore, that there are great differences among teachers, not only in the actual marks they assign but also in the qualities they intend their marks to measure. A teacher's scheme for assigning school marks becomes, in fact, so important a part of her professional character, that great opposition is frequently encountered by the principal or superintendent who undertakes to secure a more uniform marking practice throughout an entire school or school system.

Although teachers frequently object when asked to follow a uniform scheme, one may occasionally find a community or school in which a more or less common variation from the average practice has developed. In a study of the marks given during one semester in five of the large high schools of northern New Jersey, for example, it was found that in two schools only 8% of the marks indicated failure, while in the third school it was 12%, in the fourth 14%, and in the fifth 27%.[1] This difference between 8% and 27% of failure marks was certainly

[1] Some of the results of this study, especially the facts regarding the failure marks, were reported by Superintendent Don. C. Bliss, through whose coöperation the marks were collected, in an article entitled "High School Failures," in *Educational Administration and Supervision*, Vol. 3, pp. 125–138. The tabulations were made under the writer's direction by Mr. J. K. Walsh and other members of a class in educational administration at Teachers College.

not due to any difference in the general characters of the high schools and was probably not caused by differences in the intellectual abilities of the pupils. More than two thousand marks were considered in each of these five high schools, so the differences in percentages of failures cannot be attributed to differences in the sizes of the

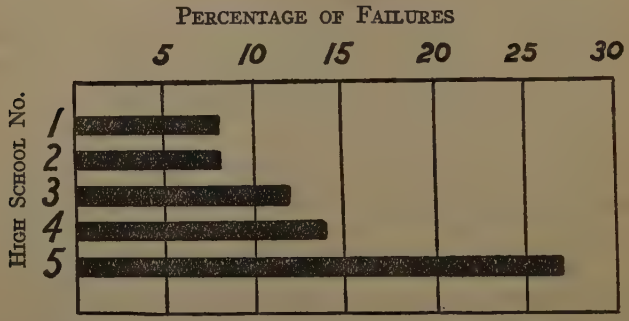


Fig. 3. — Differences between five New Jersey high schools in assigning the grade of F. Note the excessive proportion in school No. 5.

schools. The cities in which these five schools are located were nearly enough alike in their social and industrial conditions that one could not blame the differences on social or industrial inequalities. The teachers in these schools had all met approximately the same requirements for certification. The only plausible reasons for a difference of 19% in the frequency of failures was that a stronger tradition had grown up among the teachers in one of these schools to the effect that " only the best pupils are able to pass the high school subjects."

Since it is generally known that a larger proportion of pupils fail in the first two years of their high school course than in the third and fourth years, it is conceivable that the

excessive number of failures in the one high school was the result of having an oversupply of first- and second-year pupils. That this was not the case may be clearly seen by examining Table I, which gives for each of the five schools the percentage of failures in each of the four high school grades.

TABLE I

PERCENTAGES OF HIGH SCHOOL FAILURES BY CLASSES

HIGH SCHOOL NUMBER	PERCENTAGE IN GRADES				TOTAL
	I	II	III	IV	
1	12	10	5	3	8
2	10	9	4	4	8
3	17	10	6	2	12
4	19	16	1	0	14
5	31	33	25	9	27

The teachers in school No. 5 were relatively severe in their assignments of failures, not only in the first- and second-year classes, but also in the third-year class, and to some extent in the fourth-year class. Schools Nos. 1 and 2 gave a larger proportion of failure marks to first- and second-year classes than to third- and fourth-year, although the percentages were not unusually high even in the first-year classes. The general tendency to allow third- and fourth-year pupils to pass without much question is shown with special clearness in the above table by school No. 4, where 1% or fewer of the upper-grade marks were failures while from 16 to 19% of the lower-class marks were unsatisfactory. The table makes it clear, however, that the extreme

differences between schools in the percentages of reported failures were not the results of differences in the ratio of first- and second-year pupils to the entire student body. Many of the pupils who received passing marks in the high schools of cities Nos. 1 and 2 would have failed had they been required to attend the high school in city No. 5.

Differences in the marks assigned by teachers of the same subject in the same high school. — Differences in the frequency with which different schools assign marks of failure are the natural result of local traditions acting on a lack of definite standards in marking. Unless there is some opportunity for discussion and agreement as to the general policy to be followed, each teacher will develop certain tendencies which will make her marks mean something entirely different from the marks of another teacher.

Four teachers of English, in one of the five high schools (No. 2) reported above, distributed their marks in this subject as follows:

TABLE II

DISTRIBUTIONS OF HIGH SCHOOL ENGLISH MARKS BY DIFFERENT TEACHERS

TEACHER NUMBER	PERCENTAGE OF ALL MARKS				
	A	B	C	D	F (Failure)
1	9	41	28	14	8
2	2	31	64	2	0
3	23	29	44	3	1
4	12	32	49	5	1

It is hardly probable that eleven times as large a proportion of pupils reciting under teacher No. 3 were as deserving of an A as were the pupils reciting under teacher No. 2, and yet eleven times as large a proportion actually received A. It is also doubtful whether eight times as large a proportion of the pupils reciting under teacher No. 1 were as truly failures as were the pupils reciting under the other

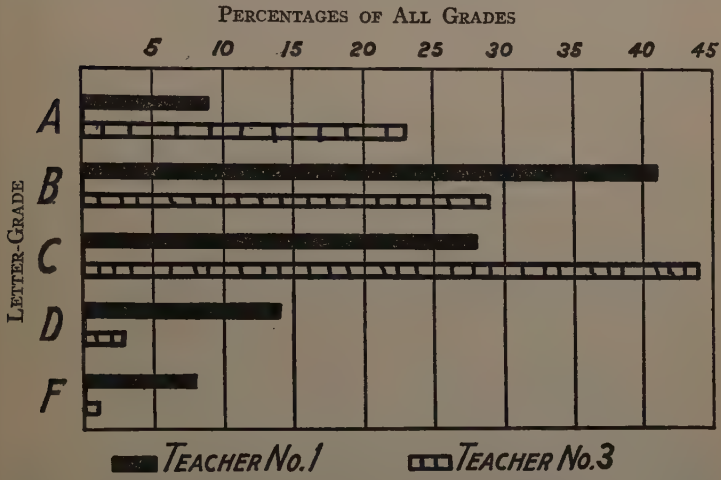


FIG. 4. — Differences in the distributions of letter-grades assigned by two English teachers in high school No. 2. Note especially the differences in the frequency of D and F marks.

teachers, although teacher No. 1 certainly failed eight times as large a proportion. Furthermore, a grade of C under teacher No. 2, of whose marks 64% were of this grade, certainly cannot mean the same as a grade of C under teacher No. 1, who gave but 28% of them. There is little doubt, however, that each of these teachers believed in the accuracy of her marks.

Three teachers of mathematics in the same high school (No. 2) distributed their grades as follows:

TABLE III

DISTRIBUTIONS OF HIGH SCHOOL MATHEMATICS MARKS BY DIFFERENT TEACHERS

TEACHER NUMBER	PERCENTAGE OF ALL MARKS				
	A	B	C	D	F (Failure)
1	6	33	27	12	21
2	31	31	17	17	4
3	5	14	39	16	26

The contrast between the distribution of marks by teacher No. 2 and that by teacher No. 3 is particularly striking, and yet no thoughtful person would care to use these percentages as evidences either of greater mathematical ability on the part of the pupils of teacher No. 2 or of poorer instruction by teacher No. 3. Either teacher No. 2 was too liberal or teacher No. 3 was too severe in the assignment of the higher marks. Certainly an A in mathematics given by teacher No. 3 meant something entirely different from an A in mathematics given by teacher No. 2.

Of the five high schools under consideration, No. 5 showed the greatest amount of variation in marks within its own ranks as well as the greatest amount of deviation from the other high schools with which it was compared. The distribution of marks given by the four teachers of English may be taken as an example of the extreme lack of uniformity found in this school:

TABLE IV

ADDITIONAL DISTRIBUTIONS OF MARKS IN ENGLISH BY DIFFERENT TEACHERS

TEACHER NUMBER	PERCENTAGE OF ALL MARKS				
	A	B	C	D	F (Failure)
1	10	55	30	5	0
2	0	0	46	29	24
3	6	24	19	20	31
4	1	29	21	29	20

It is practically impossible to believe that one teacher (No. 2) had pupils who were so poor in English that 53% of them really deserved either D or F, while another teacher (No. 1) of English in the same high school worked with pupils of such high quality that only 5% truly deserved marks as low as D. It is not probable that English teacher No. 2 obtained a group of pupils who were so inferior as to deserve no A or B marks, while English teacher No. 1 received a group so superior that 65% were worthy of A's and B's. Such wide differences in the marks assigned by teachers of the same subject in the same high school are to be explained, not by differences in the actual accomplishments of the pupils, but by differences in the standards used by the teachers in making their judgments.

Differences are the fault of the rating system. — The fact that two teachers do not agree in their assignments of school marks is in no sense a reflection on the character or ability of the teachers. The frequency and magnitude

25107

of such differences form, however, a serious criticism of the system of assigning scholarship marks. To the teacher who assigns it, a given mark has at the time a very definite significance. The teacher who rates a pupil A would usually be quite unwilling to acknowledge that the pupil deserves only a C, and by her standards she would probably be correct.

The teacher not only believes in the justice of her marks, but she carries in her mind while assigning grades a more or less definite series of mental pictures, representing the various qualities or degrees of scholarship that correspond with each mark. A series of such mental pictures might be employed as a scale for general use if it were only possible to transfer the series to the minds of other teachers. Such a transfer is not possible, however, except through a very elaborate description or representation of each picture. Even if it were possible for the originator to describe in great detail the accomplishments and merits of each individual in such a mental scale or series, an outside reader or listener would not obtain from the descriptions exactly the same understanding of the examples on the scale as was possessed by the teacher in whose mind they originated.

In so far as it exists only in the mind of the originator, a series of standards is said to be "subjective." In so far as it is made to exist in printed words, pictures, or other permanent exhibits, which can be examined and studied by other judges, such standards become "objective" in nature. It is quite clear that two judges of scholarship, or of any other trait, may not be expected to agree in their judgments except as they have a common objective standard with which to compare the new cases they are

called on to classify. As long as teachers are not given a common objective basis on which to judge the qualities of their pupils, each teacher will make up, through practice and experience, a series of subjective standards of her own. Agreement of judgments may occur as a result of similar experiences or through the interchange of ideas and the

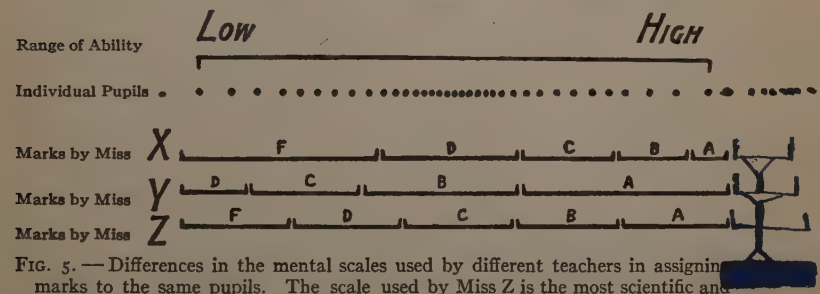


FIG. 5. — Differences in the mental scales used by different teachers in assigning marks to the same pupils. The scale used by Miss Z is the most scientific and equitable.

comparison of results, but such comparisons and common experiences are not frequent.

Classroom examinations as a basis for school marks. — May the classroom examinations conducted by the teacher serve as an adequate basis for school marks? To answer this question intelligently, one must first review the typical procedure in marking such examinations. At the end of an examination period, the answers written by the pupils are collected, and more or less carefully read by the teacher. The marks assigned to these examination papers are determined in different ways by different teachers. Some teachers read each answer carefully and rate it by assigning to it a certain number of points out of a possible ten or twenty allowed for a correct answer to that question, making the number of points allowed correspond as well as they can

to the adequacy of the answer. The score for the paper as a whole is then obtained by adding the scores made on the several questions. Other teachers read each pupil's paper as a whole and give it a general rating, according to their impression of the fitness of the writer, according to the emphasis placed by the writer on certain favorite points which were emphasized by the teacher in the class discussions, or according to some other marking plan which the particular teacher may have adopted.

Teachers differ from each other not only in their schemes for assigning grades to the papers of their pupils but also in the qualities they hope to test by the questions used in examining their pupils. Some teachers do not attempt to test abilities more complex than those involved in remembering simple facts and performing relatively simple mechanical operations. Other teachers try to frame their examination questions in such a manner as to involve a large mass of facts in the organization of answers for each question. In other words, some instructors ask only " memory questions," others ask only " thought questions," and still others attempt to build each examination so that it includes both types of questions. Among those teachers who use both types of examination questions, some insist on having a large majority of the questions depend chiefly on knowledge of definite facts, while others make the larger part of their examinations depend on ability to use or to organize facts for the solution of definite problems. There is thus no more agreement among teachers as to the kinds of questions that should be asked in examinations than there is regarding the standards by which the answers written should be judged and scored.

Differences between marks assigned the same paper by different teachers. — The classic examples of teachers' disagreements in judging the same examination papers are from the experiments conducted at the University of Wisconsin in 1912 and 1913 by Professors Starch and Elliott.[1] Two English papers were selected from those written at the end of the first year of work in one of the large high schools of Wisconsin. Plates were made of both papers in order that facsimiles might be printed in as large numbers as necessary. The set of questions and a copy of each paper were then sent to approximately two hundred high schools, " with the request that the principal teacher of first-year English grade these two papers according to the practices and standards of the school."

The questions on which the pupils had written were as follows:

1. Give five rules for the use of the comma, two for the period, and illustrate each use by a sentence.
2. Give five requirements to be observed in the structure of a paragraph.
3. Write a brief business letter.
4. Define narration, coherence, unity; classify sentences rhetorically and grammatically. Illustrate or define.
5. Name all the masterpieces studied this year and name the author of each. (Only five different selections had been studied.)
6–10. Write a three-paragraph essay, narrative, descriptive, or both combined.

[1] Daniel Starch and Edward C. Elliott: "The Reliability of Grading High School Work in English," *School Review*, Vol. 20, pp. 442–457; "The Reliability of Grading Work in Mathematics," *School Review*, Vol. 21, pp. 254–259; "The Reliability of Grading Work in History," *School Review*, Vol. 21, pp. 676–681.

Replies were received from approximately a hundred and fifty teachers, ninety-one of whom were in high schools where the passing grade was 75%. Figures 6 and 7 show

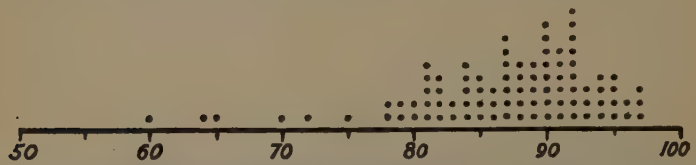


FIG. 6. — Frequency with which each percentage grade was assigned to English paper A by 91 English teachers in high schools having 75% as the passing grade. (After Starch and Elliott.)

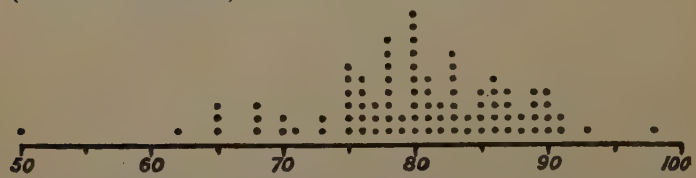


FIG. 7. — Frequency with which each percentage grade was assigned to English paper B by 91 English teachers in high schools having 75% as the passing grade. (After Starch and Elliott.)

the extreme differences of opinion among ninety-one teachers of first-year English, even when exactly the same paper was being considered, and when all teachers were using exactly the same mark as a passing grade. The percentage grades are shown below the base line, while the frequency with which each percentage grade was assigned is shown by the number of dots above that particular part of the scale.

COMPARATIVE SUMMARY OF PERCENTAGE GRADES ASSIGNED ENGLISH PAPERS

	PAPER A	PAPER B
Middle-most grade	88	80
Extreme limits	60–97	50–98
Extreme range of grades	37	48
Limits of middle half	83–92	76–86
Range of middle half	9	10

Such differences between competent teachers of the same subject are not at all unusual. Similar experiments carried on in high schools and colleges elsewhere have resulted in even wider disagreements. The Wisconsin study is mentioned here because it was one of the first, and because it had a wide influence. The corresponding experiment with the marking of a history examination paper by high school history teachers gave the distribution of marks shown in Figure 8.

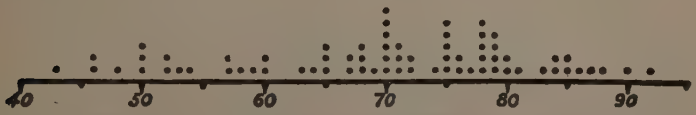


FIG. 8. — Frequency with which each percentage grade was assigned to the same history paper by 70 history teachers in high schools having 75% as the passing grade. (After Starch and Elliott.)

SUMMARY OF PERCENTAGE GRADES ASSIGNED HISTORY PAPER

Middle-most grade	71
Extreme limits	43–92
Extreme range	49
Limits of middle half	63–78
Range of middle half	15

Summary. — Scholarship marks are the most common records and reports of measurements of educational achievement. Yet such marks do not furnish reliable data for comparing the ability of a pupil at present with his ability last year, for comparing the ability of one pupil or group of pupils with another pupil or group, for the educational and vocational guidance of individual pupils, or for any other of the numerous uses to which such marks and grades are applied. There is little reason to expect teachers to have the same methods and rules in assigning scholarship

marks, for normal schools and colleges give almost no attention to this phase of the training of teachers. In actual practice teachers disagree decidedly in the way they distribute their marks, even in the same departments of the same schools. In marking exactly the same examination papers, different teachers vary as much as forty percentage points in the marks they assign. Unless teachers are taught to use an improved type of examination, a more uniform method of scoring their papers, and more objective evidences as the bases for their grades, one need not expect scholarship marks to have highly significant meanings.[1]

QUESTIONS AND EXERCISES

1. If at the end of his first month in a new school, a child should bring home a report card showing a grade of B in reading, what should his parents do about it? What would it mean to them?

2. From the story reported on pages 35 and 36, what bases were probably used by each of the teachers in the assignment of the boy's writing marks?

3. Assuming that as a teacher you wish, through the grades you assign, to secure the coöperation of your pupils' parents, would it be better to assign high grades or low grades?

4. Assuming that you as a teacher wish to stimulate your pupils to higher efforts through the grades you assign them, would you assign high grades or low grades? Why?

5. Make a list of the different considerations that have influenced you in making up the grades of individual pupils. Which of these points are really significant of ability or scholarship?

[1] Chapter XXI of this volume discusses the improvement of examinations and the introduction of a maximum amount of significance into marks and reports of scholarship.

6. Knowing that a given pupil made a grade of B in spelling last year and a grade of C this year, what facts can you state regarding his ability in spelling this year as compared with last?

7. Knowing only that Harry made a grade of B in grammar, and that James made a grade of D in the same subject, what facts can you give as to the relative abilities of these boys in grammar?

8. Examine the last set of examination questions you wrote. Just what qualities or abilities did you intend each question to measure? What proportion of other instructors in your subject would have written questions intended to measure exactly the same qualities?

9. Is it possible to classify each of the following as either a "memory question" or a "thought question"?

a. Define a right angle.
b. How many feet in a mile?
c. What is the chief industry of Detroit?
d. Why has New York City developed so extensively?
e. Who was the third president of the United States?
f. Why did the United States not join the League of Nations?

10. Hunt up an old set of examination books or papers that you graded several months ago, and have some one go through them, copying for you the list of names and the grade you previously assigned to each, and blotting out your previous marks. Grade each paper again. How much difference is there for each paper between the first grade you assigned and the second?

11. Arrange with another teacher of your subject to grade a set of your papers for you, after you have carefully recorded your estimate of each on a list which you retain. What is the maximum amount of difference between your judgment and that of the other teacher on any paper? Do you differ just as much about the best papers as you do about the poorest papers?

12. What reasons are there to expect that a given grade will mean the same degree of scholarship to all teachers who assign it?

REFERENCES FOR READING

R. E. Carter, "Correlation of Elementary Schools and High Schools," *Elementary School Teacher*, 12 : 109–118.

W. F. Dearborn, *School and University Grades*, University of Wisconsin, Bulletin No. 368.

F. W. Johnson, "A Comparative Study of Grades of Pupils from Different Elementary Schools, in Subjects of the First-Year High School," *Elementary School Teacher*, 11 : 63–68.

F. J. Kelly, *Teachers' Marks: Their Variability and Standardization.*

W. R. Miles, "Comparison of Elementary and High School Grades," *University of Iowa Studies in Education*, Vol. I, No. 1.

D. Starch, *Educational Measurements*, Chapter II.

1. Ayres - Spelling Test.
Origin, development and feature
out in the nine points opposite.
Papers - Elementary School Subjects
Thorndyke - Writing Scale
Have no items to be turned in in
textbook paper.

CHAPTER III

HOW STANDARD TESTS DEVELOPED

STANDARD tests are not the invention of any one man or group of men. Measurements in the field of education are the result of a great many independent forces acting from different directions upon the thought and practice of workers in psychology and education. The fundamental ideas, which, when combined and put to work, produced our educational scales and tests of the present day, are not entirely new, although their significance and usefulness are just beginning to be recognized. The conservatism of teachers was attacked by groups of citizens outside the schools, which suggested to certain educators a method of investigation (and of possible defense) that had been developed by the experimental psychologists. In this field as in others, " necessity " has been " the mother of invention."

The conservatism of teachers. — Teachers are employed by the state to instruct children, because it has been found more economical and usually more efficient to have children taught in groups by trained teachers than to have them taught at home by their parents. Among those who choose teaching as a calling there is a marked tendency to work things out logically and more or less schematically. It is easier to teach facts in their relations to each other than in isolation, and teachers tend to organize facts into groups

for learning purposes. The subject matter for school work is selected from the world outside the schoolroom, — from the life and experience of the race. The teacher in the school, with her tendency to organize materials, takes the experiences which are found to be helpful or instructive and weaves them into a course of study, or possibly into a textbook for the training of young minds. As soon as the course of study or textbook has been completed and published, hundreds of other teachers begin to use it as a guide for their teaching.

There is an unfortunate tendency among teachers, as well as among most other people, to reverence printed or formulated matter. It is said that in certain parts of China this amounts almost to a religious belief in the sacredness of writing, and that it is so strong that people will not destroy even a scrap of waste paper if something has been written or printed on it. In American education this tendency appears as a conservative tendency to teach just what is given in the textbook.

Teachers who have written textbooks feel a very natural pride in their authorship, which makes them slow to acknowledge that their books may now be out of date. Less original teachers frequently fail to recognize that the real materials for their work are to be found in life rather than in the limited number and variety of experiences included in the course of study. The pupils are usually not aware that they are expected to study anything broader or more vital than the textbooks placed in their hands. As a result, the curriculum of a school tends constantly to become formal and to be slow in responding to changes in life outside.

Outside influences bring pressure on the schools. — Teachers who write courses of study and textbooks seldom go out into life in active search of the experiences which will be most helpful to their pupils. As a matter of fact, new subjects have come into the schools more frequently because of pressure from groups of citizens outside than because of any search on the part of teachers for suitable subject matter. Outsiders looking on have usually been the first to discover that the subjects being taught in the schools were not preparing the pupils as directly as other studies might prepare them for intelligent participation in citizenship. This has probably been due to the fact that the teachers were thinking of the logic and arrangement of their work, while the outsiders were thinking more directly of its purposes and results.

Somewhat more than a quarter of a century ago there arose a strong and popular criticism which pointed out that the schools were devoting themselves almost exclusively to the training of the minds of their pupils and were neglecting their "hands." Reading, writing, spelling, arithmetic, geography, and history were not condemned in this criticism, but it was claimed that the pupils should have their bodies trained by the schools just as certainly as their minds. Instead of devoting itself exclusively to "books," it was suggested that the school teach its pupils how to use their hands and some tools. The "practical arts," including manual training for boys, and sewing and cooking for girls, were proposed as desirable subjects for school instruction. It was suggested that less time be given to the older subjects, and that the time thus saved be devoted to training in more practical subjects.

Early methods of discovering truth in educational work.— As one might have expected, there was much objection on the part of school people to the introduction of the new subjects, which were so radically different from the traditional work of schools. The objection was made that if less time were given to reading, spelling, arithmetic, and the other fundamental subjects, pupils would leave the schools much less well trained than otherwise. The statement was made repeatedly by those who objected to the new studies that more rather than less time ought to be given to the fundamentals.

Fortunately, a number of educators believed in the practical training which had been proposed. Their replies to the conservative teachers were not, however, more convincing than the arguments against the new subjects had been. They merely replied that *possibly* the training of the body of a child would have a beneficial effect upon his mind, so that he would perhaps learn as much of the fundamental subjects in a short time as he had previously learned in a longer time. This possibility was given very little consideration by the enemies of the practical subjects, who insisted that the time available in the school was not sufficient for teaching the usual " formal " subjects effectively, and that to give any less time to them would be an educational " crime."

Each side in the controversy began to marshal its forces and seek support from as many " authorities " as possible. It seems that in those days educational questions were decided, to a considerable extent, in favor of the side which could show the largest number of favorable opinions from " experts." Those who quoted from these " experts "

were not always careful, however, to make certain that the authority quoted was really competent to express a judgment on the particular subject under discussion. There is, of course, a real value in the judgments of intelligent persons who have had extensive experience with the particular procedure under debate, but in the controversy with regard to the comparative usefulness of practical subjects and formal subjects in the public schools, neither side could claim any extensive experience or expertness in the practical work.

A new method of determining educational facts proposed. — In the midst of this discussion it occurred to one member of the teaching profession that it would be a useful thing to *compare the results* from schools giving a great deal of time to certain of the fundamental subjects with the results from schools giving much less time. He proposed that, instead of collecting opinions as to whether giving less time to the teaching of spelling would injure the abilities of pupils to spell, it would be worth while to discover whether a school that gave little time to spelling was actually producing poorer spellers than the schools that gave more time to this subject. He suggested that exactly the same test be given in several cities in which different amounts of time had been allowed for the teaching of spelling.

"Whether or not the spelling in a particular locality is actually below the average can be learned only by comparing the results of an examination conducted on the same basis in many localities. By examining children in any one city, on a set of arbitrarily selected words, the question cannot be solved, because the results in other places, on the same list of words, would remain an unknown quantity." [1]

[1] J. M. Rice, "The Futility of the Spelling Grind," *The Forum*, Vol. 23, pp. 163–172, and 409–429.

Dr. Rice carried his plan into execution and prepared a list of fifty words which he gave as a test to pupils in certain grades as he went from town to town. In addition to the results of his test he kept a record of the amount of time allowed for the subject of spelling in each school. His early findings by this method indicated that, in so far as spelling was concerned, it made very little difference in the final results whether forty minutes or only fifteen minutes were allowed each day for instruction and drill. Fifteen minutes of spelling drill each day for eight years did not seem to produce a lower efficiency in his spelling test than was produced in other schools where more than twice that much time per day had been used for eight years.

This early effort to settle a debated question in education by means of an experimental test aroused much discussion. Teachers were so unprepared for the results found that they felt it necessary to question either the validity of the method or the honesty of Dr. Rice. It was so " absurd " to suggest that more time spent in teaching did not result in more learning on the part of pupils, that many teachers attacked the very idea of trying to compare the results of instruction in spelling by comparing the abilities of pupils to spell.

Dr. Rice seems to have been for a time almost alone in his belief in the method of giving the same test to different schools. His own faith in the possibilities of the method appears, however, to have been strengthened by his experiments. He wrote:

"The first step toward placing elementary education on a scientific basis must necessarily lie in determining what results may reasonably be expected at the end of a given period of instruction. If we

have no definite notions in regard to what our teachers ought to accomplish, our ideas must be doubly vague as to how much time need be devoted to each branch. And, as long as this remains unanswered, no well-founded opinion can be given concerning the possibility of broadening the course of study without detriment to the formal branches, — the point around which the entire question of educational reform revolves."

Another controversy yields to the new method. — There are in every line of endeavor some persons who do not look with favor on new methods and plans. Those who like to tell how much more satisfying and effective instruction was " in the good old days " will probably be with us for many years to come, although their number is becoming smaller every year. One of the strongest blows at this doctrine of the inferiority of modern education to that of a half century ago was dealt by an investigation that was essentially the same in method as that employed by Dr. Rice.

A considerable amount of criticism of the schools in New England, about two decades ago, was of this anti-modern-education variety. The introduction of new and more practical subjects had gone so far that certain people began to complain that the pupils were not mastering the older and more important subjects as well as they had " when we went to school." Business men complained that the pupils who graduated and applied to them for work could not write, calculate, and spell as well as those who had come many years earlier, and newspapers in many parts of the country took up the same complaint against the " modern fads " in education.

Most of the teachers felt that the criticism was unjust,

and many of them gave voice to their beliefs that the pupils of to-day were able to read, write, spell, and figure just as effectively as had the pupils of previous decades. They had no more proof of their assertions, however, than the critics of the schools had for their criticisms. The school masters declared, "Pupils *are* just as well taught to-day as they were fifty years ago"; and the critics replied, "Pupils are *not* as capable as they were fifty years ago."

In the *Springfield* (Mass.) *Republican* for November 12, 1905, there appeared an article by John L. Riley, at that time principal of the Central Street Grammar School in Springfield, entitled, "Schools of Sixty Years Ago." A few years previous to the appearance of this article, several old sets of examination papers had been found in the attic of the high school building at Springfield and had been preserved by the superintendent of schools in his safe. These papers consisted of written tests in spelling and penmanship and printed questions in geography and arithmetic, with the pupils' answers written on the printed sheets. It occurred to Mr. Riley that here was a chance to compare the abilities of modern pupils with the abilities of their grandparents, using exactly the same tests as measures. The article in the *Springfield Republican*[1] was the report of Mr. Riley's comparisons and gave a very conclusive answer to the question as to whether modern pupils are more or less efficient than pupils in the schools of "long ago."

The Springfield Tests: schools of the past compared with those of the present. — The spelling test consisted of

[1] Mr. Riley later published a more extensive report entitled, "The Springfield Tests, 1846–1906, A Study in the Three R's."

twenty words for the pupils to spell from the teacher's dictation, each word being written opposite its number in a previously prepared column. The words used were:

1. accidental
2. accessible
3. baptism
4. chirography
5. characteristic
6. deceitfully
7. descendant
8. eccentric
9. evanescent
10. fierceness
11. feignedly
12. ghastliness
13. gnawed
14. heiress
15. hysterics
16. imbecility
17. inconceivable
18. inconvenience
19. inefficient
20. irresistible

Few modern teachers would be willing to have their work judged by the success of their ninth-grade pupils in spelling the above list of words. It is interesting to note that although the test would be called unfair to ninth-grade pupils by the majority of present-day teachers, 245 ninth-grade pupils in 1905 had an average of 51.2% words correct, while 85 ninth-grade pupils in 1846 had obtained an average of only 40.6% correct.

As a test of penmanship the pupils were required to copy a stanza containing ten lines of poetry. One must, of course, make some allowance for the differences in the styles of writing in 1846 and 1906, but direct comparison of the samples of modern writing with the samples of the writing done by pupils of the same grade a half century ago shows a decided superiority for the modern pupils' efforts. In legibility, beauty, smoothness of line, form of letters, spacing, and every other quality, the Springfield pupils of 1906 surpassed their grandparents, or at least made a much better

showing than the pupils of the same grade in the Springfield schools in 1846.

The questions in geography and arithmetic were printed, with a space left after each question in which the child was to write his answer. The geography questions were as follows:

1. What is Latitude?
2. What is Longitude?
3. In what direction does the St. Lawrence flow?
4. Name the four large lakes between the United States and the British possessions.
5. Name the States bordering on these four lakes, and their capitals.
6. What large bay lies east of Massachusetts?
7. Name the New England States and their capitals.
8. What river is the boundary between South Carolina and Georgia?
9. Name the three large branches of the Mississippi on the west side.
10. What is the largest lake lying wholly in the United States?
11. Name the counties in Massachusetts.
12. Name the largest river in the eastern part of Massachusetts.

In spite of the narrow, old-fashioned character of the above questions, the 216 pupils who took the test in 1906 made an average of 53.4% correct, while the 81 pupils who had taken the test in 1846 made an average of only 40.3% correct. The papers of both groups of pupils were very carefully marked by the same teacher following exactly the same rules in marking both sets.

Similarly, the 79 pupils who took the arithmetic test in 1846 made an average of 29.4% correct, while the 245 pupils of the same grade who took the test in 1905 made an average of 65.5%.

Although these tests and their results are interesting evidences that the pupils of the present day accomplish as much or more in their formal studies than did the pupils of a half century ago, the chief purpose of including them here is to suggest the advantage of giving the same test over again under different circumstances. Dr. Rice, by giving the same test in schools devoting different amounts of time to spelling, was able to compare the results and to give a definite statement regarding the amount of time necessary to learn to be a speller of average ability. Mr. Riley, by giving the same tests to Springfield children in 1906 that had been given to pupils of the same grade in 1846, was able to answer definitely the criticism, which had been made without sufficient evidence, that modern pupils were not as efficient in the fundamental subjects as pupils had been " in the good old days."

Earlier attempts at uniform tests. — Although Dr. Rice has been called the " inventor " of comparative educational tests in America, it cannot be truthfully claimed that he was the first to measure more than one class with the same test. There can be no doubt that many teachers before Dr. Rice's experiments, and many since that time, found it convenient to employ an old set of examination questions rather than to make out an entirely new set. A considerable number of principals, superintendents, and state school officers had already hit upon the plan of equalizing the difficulty of final examinations by issuing a uniform test to be used in all the classes under their supervision.

Although uniformity and repetition of examination questions were not entirely unknown, few persons previous to Dr. Rice had seen the enormous advantages of such a

procedure as a basis for comparing the results secured by different teachers, different methods, and different schools. Perhaps an occasional teacher had recognized the difficulty of carrying in her mind from year to year the standards required in different grades and subjects, and had kept objective evidences of the quality of the work to which she had assigned a given mark. A few teachers had kept actual samples of their pupils' work to serve as guides in their future judgments of the same type of work. Keeping such samples of graded handwritings, compositions, or drawings assisted these teachers in giving scholarship marks that indicated the same degrees of merit with one class as with any other class in any other year.

An exceedingly interesting effort at uniformity in educational tests, and the earliest of which we have any adequate report in modern times, developed in the Greenwich Hospital School in England under the Rev. George Fisher. Although Rev. Fisher's attempts seem not to have made any permanent impression on his contemporaries, he had a clear conception of the need for objective standards in education, and his plan is quite in harmony with the best practice of the present day. The scheme was reported in 1864 by E. B. Chadwick, in the *Museum, a Quarterly Magazine of Education, Literature, and Science*, Volume III.[1]

"A book called the 'Scale-Book' has been established, which contains the numbers assigned to each degree of proficiency in the various subjects of examination: for instance, if it be required to determine the numerical equivalent corresponding to any specimen of

[1] See also *Journal of Educational Psychology*, Vol. 4, p. 551, and the *Seventeenth Yearbook of the National Society for the Study of Education*, Part II, pp. 9 and 10.

'writing' a comparison is made with the various standard specimens, which are arranged in this book in order of merit; the highest being represented by the number 1, and the lowest by 5, and the intermediate values by affixing to these numbers the fractions $\frac{1}{4}$, $\frac{1}{2}$, or $\frac{3}{4}$. So long as these standard specimens are preserved in the institution, so long will instant numerical values for proficiency in 'writing' be maintained. And since fac-similes can be multiplied without limit, the principle might be generally adopted.

"The numerical values for 'spelling' follow the same order and are made to depend upon the percentage of mistakes in writing from dictation sentences from works selected for the purpose, examples of which are contained in the 'Scale-Book,' in order to preserve the same standard of difficulty.

"By similar process values are assigned for proficiency in mathematics, navigation, Scripture, knowledge, grammar and composition, French, general history, drawing and practical science, respectively. Questions in each of these subjects are contained in the 'Scale-Book,' to serve as types, not only of difficulty, but of the nature of the questions, for the sake of future reference; observing that the same numerals are used in the same order as before, viz., number 1 denotes the highest, and number 5 the lowest amount of attainment."

Pioneer investigations with uniform tests. — Although the report of Fisher's Scale Book in England does not seem to have made much impression on the educational workers of the time, Dr. Rice's report on "The Futility of the Spelling Grind" received a great deal of public notice and criticism. Later, in 1902, Dr. Rice published a study of the arithmetical abilities of approximately six thousand school children in seventeen different schools.[1] During the same year, Dr. Cornman reported an intensive study of

[1] J. M. Rice, "Educational Research: A Test in Arithmetic," *The Forum*, Vol. 34, pp. 281–297, and "Causes of Success and Failure in Arithmetic," *The Forum*, Vol. 34, pp. 437–452.

spelling ability in the Philadelphia, Pa., public schools,[1] using some of Dr. Rice's tests and others prepared by himself for this particular investigation. Each of these studies met with severe criticism from some teachers, but each served at the same time to call the attention of thoughtful persons to the possibilities of obtaining more exact knowledge about educational results by means of uniform tests.

One of those most deeply impressed with the fundamental worth of the uniform educational test was Dr. Edward L. Thorndike, Professor of Educational Psychology at Teachers College, Columbia University. His remarkably gifted mind, thoroughly trained in psychological method and analysis, foresaw the tremendous possibilities of such tests and realized the vital necessity of safeguarding them by absolute uniformity of procedure and by careful statistical treatment and interpretation of their results. A large proportion of the credit for the present health and vitality of the educational measurements movement belongs to Professor Thorndike, under whose guidance and inspiration much of the best creative work in this field has been done. His *Introduction to the Theory of Mental and Social Measurements*, published by Teachers College in 1904, made available to students of education the essential statistical methods necessary for the scientific interpretation of educational observations and experiments, and is even now one of the most widely used books of reference in this field.

Among the "standard tests" now in general use, the Stone Reasoning Test in Arithmetic was one of the earliest

[1] O. P. Cornman, *Spelling in the Elementary School: An Experimental and Statistical Investigation.*

to be published.[1] In his report, published in 1908, Dr. Stone gave not merely the average scores made by the sixth grades in each of twenty-six school systems, but also a statistical statement of the variability of the results found in each system. In this report attention was called to some of the defects of previous studies, including the fallacy of assuming that all problems in a test were worthy of the same amount of credit. An effort was made to assign to each problem the relative weight that it should have in making up the final score of a given pupil. For more than ten years Dr. Stone's Reasoning Tests [2] remained the most highly standardized tests available for measuring the abilities of pupils to solve reasoning problems in arithmetic.

The first standard educational scales. — At the Boston meeting of the American Association for the Advancement of Science, in December, 1909, Professor Thorndike presented before Section L (Education) a scale for the measurement of merit in children's handwritings. This scale, which was published in March, 1910,[3] was the first of the scientifically constructed instruments for measuring the general quality of children's educational products. It consists of reproductions of many samples of children's handwritings arranged in groups, each group having samples of various types of handwriting of the same general merit,

[1] Cliff W. Stone, *Arithmetical Abilities and Some Factors Determining Them*, Teachers College, Columbia University, Contributions to Education, No. 19.

[2] The tests in the fundamental operations of arithmetic devised by Dr. Stone were not used extensively, probably because Courtis promptly developed better instruments for measuring ability in this field.

[3] *Teachers College Record*, Vol. II, No. 2.

and each group being just one unit better than the previous group. The difference between groups or steps on the scale are equal, — equal at least in the sense of being called " equal " by a considerable number of competent judges.[1]

In publishing this scale for handwriting, Dr. Thorndike called attention to the advantages of such scales in making comparisons of results, not merely between an individual pupil and his fellows, but more especially between his record now and his own record at some future date. With such a scale as the common basis for judgments, two samples of handwriting do not need to be compared directly with each other in order to determine which is better. Where judgments are made honestly by the scale, one may compare the records of different individuals, the records of one class with those of another, the results of one teaching method with the results of another, and the work of one school with the work of another or with its own work at some other time.

Two years after the publication of Thorndike's Handwriting Scale, there appeared a scale for the measurement of quality in English compositions.[2] The Hillegas Scale for English composition consists of ten samples of composition arranged in order of merit. The differences between steps on the Hillegas Composition Scale are not as nearly equal as are the differences in the Thorndike Handwriting Scale. The amount of difference between any two adjoining samples on the scale for composition is quite accurately determined, however, and the scale may be used with just

[1] See p. 462 ff.

[2] M. B. Hillegas, "Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record*, XIII, No. 4.

as great accuracy, though with somewhat less convenience than would have been possible had the steps been equal.

Both the handwriting scale and the composition scale consist of reproduced samples of children's school work arranged in order of general merit and are used as standards with which to compare the merit of other productions of the same sort. In general the method of making a measurement is to compare the sample to be measured with the samples on the printed scale, and to assign to the pupil's work the numerical value associated with the printed sample having most nearly the same degree of general merit. These scales measure general quality in a product, regardless of the methods used to secure the product. The basis for numbering one sample on the scale higher than another is the fact that competent teachers have judged that it is better in its quality.

Shortly after the appearance of the Hillegas Scale for English composition a scale for spelling ability was published.[1] The spelling scale consisted of words arranged according to their degrees of difficulty, as determined by the actual successes of children who tried to spell them. Instead of having the child produce some object to be compared with a set of graduated standards, the child was measured by the distance he could go in spelling more and more difficult words. The only use of judgment on the part of the examiner was in deciding whether the pupil had spelled a word correctly or wrongly.

The foregoing scales, pioneers in their respective fields,

[1] B. R. Buckingham, *Spelling Ability — Its Measurement and Distribution*, Teachers College, Columbia University, Contributions to Education, No. 59.

were all prepared under the direction of Professor Thorndike. Dr. Leonard P. Ayres of the Russell Sage Foundation was another of those who were quick to recognize the value of scientific measuring instruments in education. Ayres prepared a scale for the quality of handwriting, which had the same general purpose as that prepared by Thorndike but was evaluated and arranged somewhat differently.[1] The Ayres Spelling Scale [2] was also a decided improvement over the previously published spelling scale. The publication of scales and tests from about 1915 onward was in the hands of a much wider circle of workers than formerly, but all found it necessary to recognize the scientific leadership and professional foresight of Professor Thorndike.

The growing popularity of standard classroom tests. — A considerable amount of recognition is due Mr. S. A. Courtis of Detroit, Michigan, for his part in the development and spread of interest in standard tests of educational results. Inspired by the work of Dr. J. M. Rice and Dr. C. W. Stone with tests of ability in arithmetic, Mr. Courtis developed a series of tests to measure the speed of pupils' work in the four fundamental operations of arithmetic, using exactly the same scales in all grades from the third up through the elementary school. The results of applying his tests to about ten thousand children in some seventy schools in ten different states were reported in 1911.[3] The attention at-

[1] *A Scale for Measuring the Quality of Handwriting of School Children*, Russell Sage Foundation.

[2] *A Measuring Scale for Ability in Spelling*, Russell Sage Foundation.

[3] S. A. Courtis, "Standard Scores in Arithmetic," *The Elementary School Teacher*, Vol. XII, No. 3, pp. 127–137.

tracted by this investigation resulted in an invitation to Mr. Courtis to participate in the inquiry being made into the efficiency of the New York City public schools. His report on the application of tests to the pupils of New York City increased the reputation of the Courtis Tests and stimulated schools everywhere to try them, in the hope of determining how their own results compared with the results obtained elsewhere.[1]

The use of the Courtis Tests in the New York school inquiry set an example to other investigators of school efficiency, and objective tests of the abilities of pupils became recognized as a fundamental part of school surveys. The survey conducted by Dr. Leonard P. Ayres at Springfield, Illinois, in 1914 included standardized tests in spelling, handwriting, and arithmetic.[2] During the same year, a survey directed by Professor George D. Strayer at Butte, Montana, included standard tests of spelling, English composition, handwriting, and arithmetic.[3]

Seeing the use made of such tests in school surveys, progressive superintendents of schools throughout the United States began to take an active interest in the abilities of their own pupils as compared with the abilities of pupils elsewhere and to apply such standard tests as they

[1] S. A. Courtis, "The Courtis Tests in Arithmetic," Section D, Subdivision I, Part II, Volume I, of *Report of Committee on School Inquiry*, City of New York, pp. 389–546.

[2] L. P. Ayres, *The Public Schools of Springfield, Illinois*, Russell Sage Foundation, Publication No. E 137.

[3] G. D. Strayer et al., *Report of a Survey of the School System of Butte, Montana.* Republished in 1916 as *Some Problems of City School Administration.*

found most available. In many school systems, some of them of only moderate size, specialists in educational tests and measurements have been employed. Bureaus of educational measurements and research have been established in a considerable number of the larger cities, with comprehensive programs for measuring pupils in many subjects by means of standard objective tests and scales. Normal schools and colleges have begun to give systematic courses in educational measurements, and almost every classroom teacher has at least a vague idea that, if she only knew how to do it, she might test her pupils to discover the extent to which her instruction has been successful.

Summary. — Standard educational tests have developed gradually and are the result of many fairly independent forces acting on educational practice. The chief contributions, as far as methods are concerned, came from the experimental psychologists. Standard tests first appeared as partial defensive reactions to the criticism of conservative groups in education, who objected to the principle of modifying schools to meet certain business and vocational demands from outside the classroom. It is doubtful whether progress would have been made at such a rapid rate in this field, had tests not been recognized by Professor Thorndike and others as potentially powerful instruments for classifying pupils and evaluating instruction.

QUESTIONS AND EXERCISES

1. What influences were responsible for the introduction of the so-called "commercial subjects" into the public schools? What groups of school people objected to these subjects, and why?

2. In what fields of thought is "authority" still the source of

knowledge and truth? Will it be possible to bring to religious controversies the statistical and experimental methods that are beginning to be used in education? What will be the chief objections raised against the use of these methods in the religious field?

3. What reasons can you suggest for the failure of earlier workers to suggest and use uniform tests as a means of comparing results in education?

4. Secure the coöperation of some other teacher who has the same grade of pupils and who teaches the same subjects. Prepare a set of questions on which both of you can agree, and submit them to your classes on the same day and with exactly the same instructions. Score the papers from both classes at the same time, in order that each set may be given marks comparable to those given the other set, and then compare the scores of your pupils with the scores of the other pupils. How does the average score of your pupils compare with the average score of the other pupils? Which set contains the best paper? The poorest paper? In what respects was the test unfair to either group of pupils?

REFERENCES FOR READING

W. S. Monroe, *Theory of Educational Measurements*, Chapter I.
Seventeenth Yearbook of the National Society for the Study of Education, Part II, Chapters I and II.
G. D. Strayer, *Brief Course in the Teaching Process*, Chapter XIX.
Wilson and Hoke, *How to Measure*. Chapter I.

CHAPTER IV

A TEST FOR THE AGES OF SCHOOL PUPILS

EVERY mother who sees an attractive child in the hands of some other woman knows that the key to friendly conversation is the question, "How old is your baby?" The first question asked regarding a child by the teacher, the minister, or the physician, is almost invariably with regard to the child's age. Age is a universal scale, climbed one step at a time, by rich and poor, by great and small. To know the age of another person is to know him well in at least one respect, for each individual either has had or expects sometime to have the experience of being exactly that same age.

The importance of accuracy and comparability in the age records of pupils is emphasized as soon as one begins to measure their abilities with standard tests. The fact that a child has a score of six on a certain test begins to take on significance only when one knows that he is ten years of age and in the fourth grade, and that the typical score for ten-year-old pupils in the fourth grade is eight rather than six. Pupils differ so much in their habits of stating ages that it is desirable, in the interests of a clear understanding of the results of educational measurements, to obtain information on this subject with all possible uniformity and simplicity.

How old is a ten-year-old pupil? — Many of the printed forms used in " standard tests " have the following blanks for the pupil to fill in either before or after the test has been given:

Name.....
Age .....Grade.....
School.....

Let us suppose there are in a fourth-grade class twenty pupils who write " ten " on the blank after the word " Age." Does this mean that all twenty pupils are of the same age? No, it is not probable that any two of the twenty are exactly equal in age, unless the group contains a pair of twins. Then how nearly equal in age should one expect ten-year-old pupils to be? Experience shows that the report given by a child as to his age depends on at least three factors other than his actual age: the child's idea of the answer desired, the purposes of the questioner as understood by the child questioned, and the habits the child has formed in previous answers to that question.

The most common method of recitation, involving questions and answers, tends to develop in pupils a desire to give the answer the teacher expects. *Punch*, the English journal of humor, recently published a story illustrating this tendency. A teacher was represented as asking one of her pupils whether the earth were round or flat. The child promptly replied, " Round." " How do you know? " was the teacher's next question. " Oh, well," said the child, " I don't care to start an argument; it's flat." Where there is any possibility of more than one correct

reply, pupils almost always prefer to give the answer they think the questioner desires.

When the desire of the pupil to please the questioner is combined with the possibility of an additional personal reward to each of those who answers in a given manner, the temptation to twist the exact facts as far as possible in the desired direction is too great for most pupils to withstand. If the principal of a school walks with a cheerful mood into a fourth-grade room and asks with a smile to see the hands of all pupils who are ten years old, he will probably find among those whose hands are raised many who are more than ten and a few who are less than ten years of age. On the other hand, if he appears in a stormy mood, demanding that all ten-year-old pupils hold up their hands, the response will usually be less generous. For similar reasons, the number of pupils who will write " ten " on the blank after the word " Age " on a test, will depend to a certain extent on the understanding the pupils have as to the purposes of the test.

Some teachers request their pupils to give their ages to the nearest birthday, while others ask for " your age at your last birthday." There is, on the other hand, a very general tendency among children to regard age as a virtue and to give ages at next birthday where the form of the question will allow it. Sometimes a pupil whose regard for truth is keen will qualify his claim to advanced age by saying that he is " going on eleven." Many pupils, however, even at the early age of ten, are not so cautious in their claims, especially where it does not seem that any one will be injured by the exaggeration.

A " ten-year-old " pupil may, therefore, actually be any-

where from 9 years, 0 days, up to 10 years, 365 days of age, depending on his ambitions, his hopes of reward, and his habits of calculation. The most frequent meaning of " ten years old " in adult usage is probably that the child has passed his tenth birthday but has not yet reached his eleventh, — that he is somewhere between 10 years, 0 days, and 10 years, 365 days. The life insurance companies, however, for statistical reasons that will appear later, calculate ages to the nearest birthday and would consider a child " ten years old " as long as he was between 9 years, 6 months, 0 days, and 10 years, 6 months, 0 days. In almost no calculation except in oriental countries would " ten years old " mean from 9 years, 0 days, up to 10 years, 0 days, although this period is properly called the child's tenth year, for the same reasons as the present century is called the twentieth century.

The practice of school administrators in making age-grade tables by half-year periods has further complicated the situation with respect to the exact meaning of " ten years old." A few superintendents have made such reports, calculating the ages to the last full half year, calling all pupils " ten years old " who were between 10 years, 0 days, and 10 years, 6 months, 0 days; and calling " ten and a half years old " those between 10 years, 6 months, 0 days, and 11 years, 0 days. It is much more common practice, however, and more useful from a statistical point of view, to calculate to the nearest half year, calling all pupils " ten years old " who are between 9 years, 9 months, 0 days, and 10 years, 3 months, 0 days; and calling " ten and a half years old " those between 10 years, 3 months, 0 days, and 10 years, 9 months, 0 days. The best authori-

ties advocate calculations to the *nearest* half year whenever it is desired to group ages by half-year periods.[1]

AGE LIMITS OF TEN-YEAR-OLDS

Calculated to	Limits
Last birthday	9 yrs., 0 days — 9 yrs., 365 days
Nearest birthday	9 yrs., 6 mos. — 10 yrs., 6 mos.
Last full half year	10 yrs., 0 mos. — 10 yrs., 6 mos.
Nearest half year	9 yrs., 9 mos. — 10 yrs., 3 mos.

How to secure a record of children's ages. — It is not the intention of this chapter to dictate just what system shall be used by any investigator in grouping the ages of pupils, but it is desired that every reader shall be keenly alive to the errors that may creep into the interpretation of the results of educational tests if ages are not accurately obtained and most carefully tabulated. The blanks on which pupils are to write their ages should call for definite and specific information. They should read, "Age at last birthday," or "Age at nearest birthday," rather than "Age" without any indication of how to determine that common but variously calculated fact. The most accurate method of determining age is, of course, to copy it from the cumulative record card maintained for each pupil by every modern school, which record gives the date of birth as indicated on the baptismal or birth certificate of the child. The next best method, and the most accurate one where it is necessary to take the pupil's own word for his age, is to ask two questions: "What was the date of your last birthday?" and "How old were you?"

[1] See pages 129–136 of Strayer and Engelhardt, *The Classroom Teacher*.

Experience has demonstrated that it is not usually wise to ask pupils to calculate their ages in years, months, and days, because of the errors they constantly make in such calculations. It is likewise not the safest policy to ask them to write their dates of birth, for this also frequently involves a calculation in which errors may easily be made. Practically every American school child, however, can give the date of his last birthday and tell how old he was at his last birthday. From this information the teacher may by means of age computation tables [1] quickly determine the child's exact age in years, months, and days. The use of age computation tables, either printed or home-made, is always advisable, even when mathematically inclined teachers are transmuting the dates of birth into exact ages or age groups. This recommendation is based on the general principle that all routine statistical operations should be made as simple and mechanical as possible, in order to avoid errors of computation, and in order to release, for the interpretation and application of the results, energies which would otherwise be used up in routine computations.

How old are fourth-grade pupils? — If the law required that children begin school when they were six years old, and if, beginning in the first grade, pupils spent one year in each school grade, one would expect them to enter the fourth grade when they were nine years old. If the school authorities interpreted the law to mean that pupils were six years old and ready to enter school when they were from 6 years, 0 days, to 6 years, 364 days of age, then one

[1] See especially the *Strayer-Engelhardt Complete Age Computation Tables;* also described on pages 131–132 of *The Classroom Teacher.*

TABLE V

Sample Table for Computing the Ages of School Children from Dates of Birth

Ages to the nearest half year on March 1, 1923

Date of Birth on or between	Children's Ages Range from	to	Age Group
Dec. 1, 1917 and May 31, 1918	4 yrs., 9 mos.	up to 5 yrs., 3 mos.	5 yrs.
June 1, 1917 " Nov. 30, 1917	5 yrs., 3 mos.	" " 5 yrs., 9 mos.	5½ "
Dec. 1, 1916 " May 31, 1917	5 yrs., 9 mos.	" " 6 yrs., 3 mos.	6 "
June 1, 1916 " Nov. 30, 1916	6 yrs., 3 mos.	" " 6 yrs., 9 mos.	6½ "
Dec. 1, 1915 " May 31, 1916	6 yrs., 9 mos.	" " 7 yrs., 3 mos.	7 "
June 1, 1915 " Nov. 30, 1915	7 yrs., 3 mos.	" " 7 yrs., 9 mos.	7½ "
Dec. 1, 1914 " May 31, 1915	7 yrs., 9 mos.	" " 8 yrs., 3 mos.	8 "
June 1, 1914 " Nov. 30, 1914	8 yrs., 3 mos.	" " 8 yrs., 9 mos.	8½ "
Dec. 1, 1913 " May 31, 1914	8 yrs., 9 mos.	" " 9 yrs., 3 mos.	9 "
June 1, 1913 " Nov. 30, 1913	9 yrs., 3 mos.	" " 9 yrs., 9 mos.	9½ "
Dec. 1, 1912 " May 31, 1913	9 yrs., 9 mos.	" " 10 yrs., 3 mos.	10 "
June 1, 1912 " Nov. 30, 1912	10 yrs., 3 mos.	" " 10 yrs., 9 mos.	10½ "
Dec. 1, 1911 " May 31, 1912	10 yrs., 9 mos.	" " 11 yrs., 3 mos.	11 "
June 1, 1911 " Nov. 30, 1911	11 yrs., 3 mos.	" " 11 yrs., 9 mos.	11½ "
Dec. 1, 1910 " May 31, 1911	11 yrs., 9 mos.	" " 12 yrs., 3 mos.	12 "
June 1, 1910 " Nov. 30, 1910	12 yrs., 3 mos.	" " 12 yrs., 9 mos.	12½ "
Dec. 1, 1909 " May 31, 1910	12 yrs., 9 mos.	" " 13 yrs., 3 mos.	13 "
June 1, 1909 " Nov. 30, 1909	13 yrs., 3 mos.	" " 13 yrs., 9 mos.	13½ "
Dec. 1, 1908 " May 31, 1909	13 yrs., 9 mos.	" " 14 yrs., 3 mos.	14 "
June 1, 1908 " Nov. 30, 1908	14 yrs., 3 mos.	" " 14 yrs., 9 mos.	14½ "
Dec. 1, 1907 " May 31, 1908	14 yrs., 9 mos.	" " 15 yrs., 3 mos.	15 "
June 1, 1907 " Nov. 30, 1907	15 yrs., 3 mos.	" " 15 yrs., 9 mos.	15½, "
Dec. 1, 1906 " May 31, 1907	15 yrs., 9 mos.	" " 16 yrs., 3 mos.	16 "
June 1, 1906 " Nov. 30, 1906	16 yrs., 3 mos.	" " 16 yrs., 9 mos.	16½ "
Dec. 1, 1905 " May 31, 1906	16 yrs., 9 mos.	" " 17 yrs., 3 mos.	17 "
June 1, 1905 " Nov. 30, 1905	17 yrs., 3 mos.	" " 17 yrs., 9 mos.	17½ "

Note. This table supplies the ages of pupils to the nearest half year on March 1, 1923. If ages to the nearest year or to the last birthday are desired, or if ages on a different date are desired, another table for computing ages must be prepared. This one is presented merely to show the arrangement and convenience of such tables. *Example:* Any pupil whose date of birth falls on or between June 1, 1913, and November 30, 1913, is somewhere between 9 years, 3 months and 9 years, 9 months of age on March 1, 1923, and may be classified as 9½ years old, if the grouping is to be made to the nearest half year.

should expect these pupils three years later to be between 9 years, 0 days, and 10 years, 0 days. Under such conditions, the average age of pupils entering the fourth grade would be approximately nine and a half years.

In actual fact pupils do not always enter the first grade when six years old. Some enter before and some enter after they are six. Pupils do not all make progress through the grades at the regular rate of one year to a grade. A few pupils complete the first three grades in less than three years, and many pupils spend more than three years in completing these grades. As a result, pupils entering the fourth grade are not all nine years of age, — there are some eight years old, many ten years old, and some eleven or twelve years of age. On account of these irregularities, the average age of pupils entering the fourth grade may not be the nine and a half years which normal entrance and progress would lead one to expect.

As an example of the ages actually found among pupils in the fourth grade, the following dates of birth, copied from the cumulative record cards of the ten pupils found in this grade in a country school on March 1, 1923, may be considered:

May 3, 1911	January 18, 1913
August 15, 1911	July 6, 1913
April 29, 1912	August 1, 1913
June 22, 1912	December 9, 1913
September 5, 1912	October 24, 1914

Comparing the dates of birth with Table V, one finds the ages of these ten pupils, calculated to the nearest half year on March 1, 1923, to be as follows: 12, $11\frac{1}{2}$, 11, $10\frac{1}{2}$, $10\frac{1}{2}$, 10, $9\frac{1}{2}$, $9\frac{1}{2}$, 9, and $8\frac{1}{2}$. The mean age of these ten

pupils would be found by adding together all ten ages and dividing by ten. If one accepts the grouping to the nearest half year and adds the ages as listed above, the sum is one hundred and two years. Dividing by ten, the number of pupils, gives ten and one fifth years as a quotient. The average or mean age of these pupils, by this calculation, is therefore 10 years, 2 months, and 12 days.

If actual ages in years, months, and days are used as the basis for calculating the mean age of a group, the result will frequently differ slightly from the results obtained when various methods of grouping ages are used. The actual ages on March 1, 1923, of the ten pupils used in the illustration would be as follows:

12 years, 2 months, 3 days
11 years, 5 months, 15 days
11 years, 1 month, 29 days
10 years, 3 months, 22 days
10 years, 6 months, 5 days
9 years, 10 months, 18 days
9 years, 4 months, 6 days
9 years, 5 months, 0 days
8 years, 9 months, 9 days
8 years, 7 months, 24 days

The sum of these more exact calculations of the ages of the ten fourth-grade pupils is 101 years, 8 months, and 11 days. Dividing this sum by ten gives 10 years, 2 months, and $1\frac{1}{10}$ days as the mean age of pupils in this class. This result differs by only a few days from that obtained in the previous paragraph. The average age of these particular fourth-grade pupils was therefore approximately 10 years and 2 months on March 1, 1923. This is eight months more than

one should expect of pupils entering the fourth grade. Assuming that these pupils entered the fourth grade six months earlier, on September 1, 1922, one may say that as far as their average age is concerned they are only about two months older than one should expect. It is quite obvious, however, that the entire list of ages gives a better idea of the ages of the pupils than can be obtained from their average score alone.

An error in the usual calculations of average age. — Ages have not usually been grouped to the nearest half year, as was done in the above discussion. It has been much more common to classify ages "to the last birthday." If this had been done in the case of the ten fourth-grade pupils, their ages would have read as follows: 12, 11, 11, 10, 10, 9, 9, 9, 8, and 8.

Many people, if called on to find the average or mean age of the ten pupils, would first calculate the age of each pupil to the last birthday, as in the above paragraph, and would then divide the sum of these age scores by their number. The sum of the age scores calculated to the last birthday is 97 years. The quotient obtained by dividing 97 years by ten is $9\frac{7}{10}$ years, which many intelligent but uncritical people would accept as the average age for this group of pupils. Calculations made earlier with ages grouped to the nearest half year, and further calculations made with the ages given exactly to the month and day, demonstrated that the mean or average age of these ten pupils was 10 years and 2 months. The grouping to the last birthday seems to give an average age which is approximately 6 months less.

The reasons for this difference in mean age, even though

exactly the same pupils are being measured for age and on the same day, are to be found in the failure of the usual calculations to recognize that a ten-year-old pupil, by the last-birthday definition, may be anywhere from 10 years, 0 days up to 10 years, 365 days of age. Exactly 10 years as a measure of his age is likely to be just as much below the truth as an even 11 years would be above it.

Figure 9 shows this situation fairly clearly. It will be observed that if ages are to be calculated to the last birth-

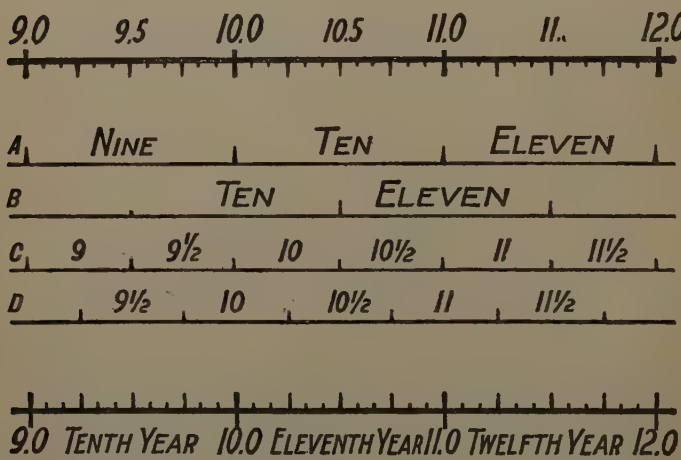

9.0 9.5 10.0 10.5 11.0 11.. 12.0
A NINE TEN ELEVEN
B TEN ELEVEN
C 9 9½ 10 10½ 11 11½
D 9½ 10 10½ 11 11½
9.0 TENTH YEAR 10.0 ELEVENTH YEAR 11.0 TWELFTH YEAR 12.0

FIG. 9.— Different ways of grouping ages: *A*. To the last full year. *B*. To the nearest year. *C*. To the last half-year. *D*. To the nearest half-year.

day, then 10 years, 6 months, 0 days (or $10\frac{1}{2}$ years) will be at the middle of the range of ages called "ten," and will better represent the actual ages of pupils called "ten years old" than any other number could represent them. It is also clear that the use of integral numbers to represent ages calculated to the last birthday will always tend to give

an average age that is incorrect by about six months. The use of 9 to represent ages from 9.0 to 9.999 years, of 10 to represent ages from 10.0 to 10.999 years, and so on, will always give an average that is .5 of a year too low.

The continuity of time and age. — The mean age of a group of persons at any given time is a definite number of years, months, days, hours, and minutes. Since no useful purpose would be served by keeping ages in minutes and seconds, however, records are kept by the schools of only the day or date of a child's birth. For most purposes a person's age in years is sufficient; hence the school practice of grouping ages by years or half years has developed. It should be remembered, however, that after such grouping has been made, it is impossible to convert the approximate age of a child back into a more exact age. If ages are known to have been stated carefully " to the last birthday," for example, one may assume that the average age of all who were called ten was approximately 10 years, 6 months; while if the grouping was " to the nearest birthday," one may safely assume that the average age of all who were called ten was 10 years, 0 months, 0 days. (See Figure 10.)

Such assumptions are possible because time is a continuous affair. For convenience in recording other facts, we divide it into days, months, years, and centuries, but there are no gaps or vacant spots in time or age. One is not the same age this instant as he was one-tenth of a second ago, — he is one-tenth of a second older. The present moment followed the one before it, which followed the one before it, which followed the one before it, and so on back to the very beginning of all time. Time moves on in

an endless stream, and as it moves on our ages are constantly growing greater and greater.

Associated with the concept of the continuity of time, there is, in some of our calculations of age, the further assumption that at no particular period in the day, in the month, or in the year, are births more likely to occur than at any other period. If we know that one hundred children were born in one month, we may assume that half of them were born before the middle of the month and that the other half were born after the middle of the month. Of a hundred children whose ages at last birthday were ten years, fifty of them may be assumed to be between 10 years, 0 months, 0 days and 10 years, 6 months, 0 days; while the other fifty may be assumed to be from 10 years, 6 months, 0 days up to 10 years, 12 months, 0 days.

In this connection it should be mentioned that the age scale is not the only one which is continuous. As a matter of fact, every scale used in education tends to be continuous, although progress along other scales is not usually as regular and even as along the age scale. Every item the child learns in his reading contributes to many other abilities. The ability to spell or to solve problems in arithmetic increases or decreases even when the pupil is studying geography or playing marbles. One does not reach a certain point in ability by an instantaneous jump and then remain at exactly that point until another jump is made. The human mind depends on muscular and nervous tissues which are constantly changing, developing, and dying. The strength of any human characteristic is, therefore, not exactly the same in an individual pupil to-day as it was yesterday. Changes are constantly going

on in every ability. The differences may be so small as not to make any change in the score on a test, but the score on the test is merely the rough grouping of abilities in the same sense as " age ten " is the rough grouping of an infinite number of differing ages. For this reason, the calculation of average or mean scores on an educational scale must take into account the limits and range of each score, just as similar calculations of average age must consider the limits and range of possible ages in each age group.

Summary. — Since educational measurements are essentially more exact descriptions of educational traits and abilities, it is not entirely satisfactory to report the results of educational measurements in terms of pupils' ages, unless these ages themselves have been so collected that their meanings are capable of definite and exact interpretation. The average score for pupils at a given age may be obtained by grouping together the scores of those who are of approximately the same age, provided the grouping is done accurately by a definite rule, and provided further that the calculation of the average takes into account the range and limits of each age group. It is not satisfactory for the purposes of educational measurements to ask the pupil merely to give his age.

QUESTIONS AND EXERCISES

1. Try the experiment described on page 82 to determine the effect of the tone and mood of the questioner on the responses of the pupils. Use the age to which the largest number of pupils belong. What other factors disturb the validity of your results in this test?

2. The names and ages of six pupils graduating from a rural school eighth-grade class were as follows:

Black, James	14 years, 8 months, 11 days
Eastman, Mary	14 years, 2 months, 25 days
Monroe, Ella	15 years, 5 months, 5 days
Norton, Harry	13 years, 11 months, 19 days
Smith, Mary	14 years, 6 months, 28 days
Taylor, Helen	15 years, 3 months, 1 day

(*a*) What was the mean age of these six pupils when calculated from the exact ages given above?

(*b*) Calculate the mean age of these pupils from their approximate ages, if the approximations were to the nearest year.

(*c*) What would be the mean age calculated from the approximate ages if these were stated to the nearest half year?

(*d*) Calculate the mean of the six ages approximated to the last birthday. What correction is necessary in this calculation that was not required in the first three calculations?

3. Prepare an age computation table, similar to that on page 86, for children from $4\frac{1}{2}$ to $18\frac{1}{2}$ years of age on September 1, 1924.

4. By means of the table you have just prepared give the ages to the nearest half year on September 1, 1924, of each of the pupils whose dates of birth are given in the following list.

DATES OF BIRTH OF SIXTH-GRADE PUPILS

6A	6B
Dec. 7, 1910	Aug. 2, 1912
Aug. 28, 1913	Nov. 29, 1912
Jan. 13, 1913	June 25, 1913
July 8, 1914	May 28, 1913
Sept. 20, 1912	June 3, 1913
May 1, 1914	F b. 7, 1914
Apr. 12, 1913	May 14, 1915
Dec. 5, 1914	Aug. 1, 1911
Nov. 1, 1912	Oct. 2, 1913
Oct. 14, 1913	Dec. 14, 1912

6A	6B
Jan. 10, 1914	Apr. 21, 1914
Nov. 24, 1913	Oct. 9, 1914
Oct. 18, 1912	Nov. 14, 1911
June 23, 1912	Apr. 13, 1914
Feb. 3, 1915	Aug. 23, 1912
Mar. 8, 1913	June 5, 1911
Jan. 6, 1909	June 16, 1912
May 26, 1914	Sept. 12, 1913
Nov. 24, 1914	June 5, 1914
July 17, 1913	Jan. 8, 1912
Mar. 2, 1914	Dec. 7, 1913
Aug. 9, 1910	Aug. 10, 1913
July 28, 1912	Dec. 27, 1911
May 12, 1914	Dec. 9, 1912
Jan. 2, 1915	May 11, 1913
Jan. 5, 1911	Feb. 4, 1913
Feb. 15, 1912	Nov. 5, 1913
July 4, 1911	Mar. 24, 1913

5. How much difference is there in the mean ages of pupils in the two sections of the class? (Make your calculations from the approximate ages found in 4.)

6. Make a list of the ages in each section approximated to the last birthday. Calculate the mean age of each section from these approximations. How do these means compare with those found in 5?

REFERENCES FOR READING

F. P. Bachman, "Over Age and Method of Determining Over Age," *Report to the New York Committee on School Inquiry*.

L. N. Hines, *Guide to Educational Measurements*, Chapter II.

Report of the Survey of the Public School System of Baltimore, Maryland, Volume 2, pp. 128–191.

Strayer and Engelhardt, *The Classroom Teacher*, Chapter VIII.

CHAPTER V

DISTRIBUTIONS OF AGE SCORES

AGE is so important a characteristic of all school pupils that it is desirable to discuss rather fully not only how records of age should be obtained but also how such records should be tabulated and used. If significant conclusions are to be drawn from the comparison of pupils' scores on tests at a given age, or from the comparison of pupils' ages when they attain a given score on a test, ages must not only be obtained with accuracy but they must also be tabulated and interpreted with uniform care and sound statistical methods.

The inadequacy of average scores. — The average or mean age of a family of two, consisting of a husband at twenty-five and a wife at twenty-three years, would be twenty-four years. If a child were to be born to them, however, the mean age of the family would at once become sixteen years (one-third of the sum of 25, 23, and 0 is 16). If another child were born two years later, the mean age of the family would then become thirteen and a half years (one-fourth of the sum of 27, 25, 2, and 0 is $13\frac{1}{2}$). To compare one family with another as to age requires a much more detailed report than a mere statement of mean ages, unless gross misunderstandings of the entire situation are to occur.

Every year, however, some "investigator" publishes the "average score" of some school measured on some educational test, and in not a few cases these average scores are compared with the average scores obtained in other schools. In one school the results used in obtaining the average may be from grades 5, 6, 7, and 8 combined, while in another the results may be for pupils in grades 3, 4, 5, and 6, but these differences in the composition of the "schools" are not always considered.

The scales for educational achievement seem to be used in such comparisons much more frequently than the scale for age, probably because the age scale is more accurately and intimately known than the scales for ability in school subjects. Although it has been fairly common to see schools compared as to average handwriting, one does not find many cases where schools have been compared as to the average age of the pupils. It is possibly just a bit easier to be sensible and critical of one's thought about the ages of pupils in different schools than to be sensible and critical of one's statements about the qualities of the compositions written by the pupils in these schools. If one attempts to find the solution of definite educational problems, however, one must be critical and must use common sense in all one's thoughts and statements, particularly when these concern the comparative status of two or more complex situations.

Comparisons of groups of individuals by their mean or average scores is never entirely satisfactory and is frequently misleading. In certain cases where the individuals composing a group are not closely alike in the characteristic being studied, the average score may actually give a

false impression of the relative standing of the one group as compared to other groups. Consider, for example, the rural school in which the eighth-grade pupils' ages to the nearest birthday on the first day of school were, in order, 12, 12, 13, 14, 15, while the seventh-grade pupils' ages were 10, 11, 12, 14, 15, 16. The mean age of the eighth-grade pupils was 13.2 years, while the mean age of the seventh-grade pupils was 13.0 years. To report no more than that the mean age in the eighth grade was .2 of a year greater than the mean age in the seventh grade would give an impression of the situation which would be not only inadequate but actually erroneous. The young pupils in the seventh grade were, for example, relatively younger for their grade than were the young pupils in the eighth, although the mean age in the seventh grade seemed to indicate greater excess age than was indicated by the mean age in the eighth grade.

Seriation as an aid to the interpretation of scores. — The age situation in the two classes described in the above paragraph is much clearer from the list of ages as given than it would have been if the age scores of the pupils had not been arranged in order according to size. If the age scores of the two classes compared had been arranged according to the alphabetical order of the pupils' names, the reader's picture of the situation would have been much less clear. Original data frequently do not mean anything to those who gather them, chiefly because they are unorganized. The mere arrangement of the results of measurement according to size is a helpful step if one has not learned to make and interpret tabulations of a more elaborate sort.

TABLE VI

AGES OF PUPILS IN TWO GRADES

ALPHABETICAL ORDER		SERIAL ORDER	
VIII	VII	VIII	VII
14	12	15	16
12	14	14	15
13	16	13	14
15	10	12	12
12	15	12	11
	11		10

The arrangement of age scores or other facts according to size may be called *seriation*. The group of scores when finally arranged in serial order according to size is frequently called an *array*.

Although seriation is not the most satisfactory method of interpreting scores for comparison with other groups of similar scores, it is sometimes an excellent device for comparing with each other the scores of the different individuals composing a single group.[1] In the eighth-grade scores of the above table, for example, the serial arrangement from oldest to youngest would be much more sig-

[1] In reporting the scores of soldiers in intelligence tests, the writer assumed at first that commanding officers would prefer to have the lists in alphabetical order; but actual trial proved that officers were able to understand and use reports on their men much more fully if the names and scores were arranged in numerical or serial order beginning with those men whose scores were highest and ending with those whose scores were lowest.

nificant and effective as a comparison of individuals than the alphabetical arrangement would be.

TABLE VII

EIGHTH-GRADE PUPILS' AGES

ALPHABETICAL ORDER		SERIAL ORDER	
Name	Score	Name	Score
Adams, Jos.	14	Thomas, Helen	15
Hoyt, Minnie	12	Adams, Jos.	14
Smith, Margaret	13	Smith, Margaret	13
Thomas, Helen	15	Wallace, Harry	12
Wallace, Harry	12	Hoyt, Minnie	12

Frequency tables reveal the significant facts regarding a group. — The significant age facts regarding the seventh- and eighth-grade groups compared above are more clearly revealed by a "tabulation" of the individual age scores, which shows how many in each class were at each age level. Such a tabulation is known as a *frequency table.* Table VIII reads from left to right as follows: "At age 16 there was one pupil in the seventh grade and none in the eighth grade; at age 15 there was one pupil in the seventh grade and one in the eighth"; and so on. It shows that the seventh grade was distributed over a wider range of ages than the eighth. Seven age groups were necessary in order to include all of the seventh grade, while only four age groups were necessary for the eighth grade. While the eighth-grade scores tended to cluster closely about their mean, those of the seventh grade showed a

TABLE VIII

AN AGE FREQUENCY TABLE

AGE GROUP[1]	NO. OF PUPILS BELONGING TO	
	Grade VII	Grade VIII
16	1	–
15	1	1
14	1	1
13	–	1
12	1	2
11	1	–
10	1	–

tendency to scatter toward the extremes. Grade VIII was, therefore, in so far as age is concerned, more compact or "homogeneous" than Grade VII.

As a further example of how frequency tables assist one in classifying the essential facts regarding groups of individual measurements, one may consider Table IX, showing the frequency of pupils of each age entering the fourth year of the high schools in two cities, A and B. Table IX is to be read as follows: "At age 23, there is 1 pupil in City A and none in City B; at age 22, there are 2 in City A and none in City B; at age 21, there are 3 in City A and 2 in City B"; and so on down to age 15, where there are 2 in City A and 9 in City B.

A glance at the two distributions in this table shows that pupils, when they enter fourth-year high school, tend to be younger in City B than in City A. By selecting any particular age, 20 for example, one can almost immediately

[1] Ages calculated to nearest birthday.

TABLE IX

DISTRIBUTION OF AGES OF FOURTH-YEAR HIGH SCHOOL PUPILS

AGE GROUP[1]	NO. OF PUPILS BELONGING IN	
	City A	City B
23	1	–
22	2	–
21	3	2
20	22	4
19	46	26
18	113	111
17	79	164
16	36	60
15	2	9
Total	304	376

make such comparisons as that City B has in the fourth year of its high school only 6 pupils of twenty years or older, while City A has 28 pupils who are twenty or more years of age. At the other extreme, City B has 233 pupils below eighteen years of age, while City A has only 117 below this age. Such comparisons are significant and useful, but they could not be made from the original unorganized age scores of the pupils without much laborious counting and checking. Distributions showing the frequency of scores of each size in each place make such comparisons and many others readily available. Without such frequency tables it is a long and tedious process to

[1] Ages calculated to last birthday on September 1. The illustrations used here give ages at last birthday, not because they are more desirable than ages at nearest birthday, but because they are now more frequently so reported.

obtain any comparisons of the two situations other than the averages, which are always inadequate and sometimes misleading bases for comparing groups of scores.

Surfaces of frequency represent groups of scores. — Tables of frequency enable one to make extremely useful comparisons between two or more groups of scores, but even with such tables there may be characteristics of the group as a whole which do not stand out for one's consideration as quickly or as forcefully as they would in a graphic surface [1] of frequency. The frequencies of Table VIII are each represented in Figure 10 by a rectangle of a

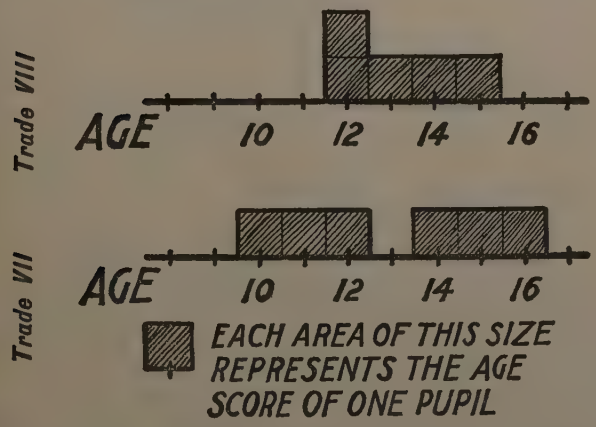


FIG. 10. — Graphic comparison of surfaces of frequency showing ages of seventh- and eighth-grade pupils. Each pupil's age is represented by a shaded square above an appropriate section of the base line. The two distributions are exactly the same as in Table VIII.

[1] A more exact but less frequently used name for such a figure is "histogram." The word "surface," when used in the phrase "surface of frequency," indicates the area of plane surface included between the irregular bounding line and the straight horizontal base line.

given size, located above the appropriate score on the base line of the figure. For the seventh grade the figure is only one unit high and extends from 10 to 16, with the exception that there is nothing at 13. For the eighth grade, however, the figure is one unit high at ages 15, 14, and 13, but is two units high at 12, because there were two 12-year old pupils in this grade. In this particular com-

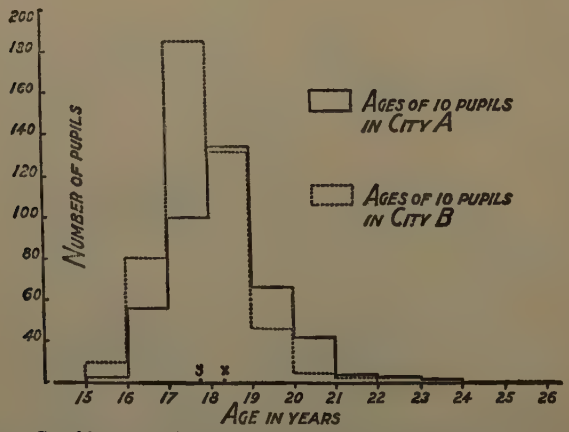


FIG. 11. — Graphic comparison of age distributions for fourth-year high school pupils in City A and City B. These surfaces of frequency show facts given in Table IX.

parison the frequency tabulation in Table VIII gave almost as complete and simple an understanding of the entire situation as is given by the surfaces of frequency in Figure 10, but this would not be true of tabulations involving large numbers of individuals.

The frequencies of Table IX are represented graphically in Figure 11 in exactly the same manner as the frequencies of Table VIII were represented in Figure 10. The surface

for City B is represented by a dotted line, and is superimposed on that for City A, in order to show more directly the differences between the two groups. The area above any section of the common base line and below the irregular boundary line shows the frequency of scores of that particular size.

It will be observed that in the older ages the frequency surface for City B is in general lower (less high above the line) than that for City A, but is higher than that for City A in the younger ages. This is merely a graphic representation of the fact that the pupils in City B are grouped on lower ranges of the age scale than are the pupils in City A. The "mode," or highest point, comes at age 18 for City A and at age 17 for City B. The mean age in City A, represented by the point x, is higher in the age scale (farther to the right on the base line) than the mean age in City B, which is represented by the point y. From the picture in Figure 11, the two situations are more easily and completely grasped by the ordinary mind than from the numbers listed in Table IX.

A modification of such surfaces of frequency is often valuable when representing to the pupils themselves their relative scores in a test. The age scores of the eighth-grade pupils tabulated in Tables VII and VIII, and shown graphically in Figure 10, are represented again in Figure 12. Each pupil's initials are entered in the rectangle that represents his score. From such a modified frequency surface, Helen Thomas would quickly see that she had the highest age score of the group, while Harry Wallace and Minnie Hoyt would be interested in finding themselves tied for the low score. If these were scores

in a test on which a pupil might by extra effort change his relative position in the group, such graphic representation of the scores would have a stimulating effect on almost all pupils. For the purposes of comparing one group with another, only the outline or boundary of the frequency surface is used. But for the teacher, who wishes to make changes in the abilities of her class, the identification of

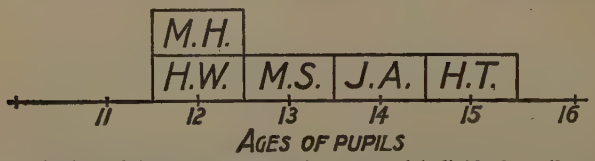


FIG. 12. — Surface of frequency representing scores of individual pupils. Initials identify the area representing each pupil's age.

each block of the surface with some known pupil is often of real assistance.

Methods of preparing frequency tables. — Distribution tables, showing the frequency of each different size of score in each group, are so much more readily interpreted than are unorganized masses of results, that scores made by the individual members of a group should always be tabulated in a frequency distribution. The common practice in making such tabulations is to write in a column all the different sizes of scores, arranged in the order of their magnitude, from large to small. The writer's students have usually found it less confusing, in their later work with such frequency tables, to have the larger or higher magnitudes at the top of the column and the smaller or lower magnitudes at the bottom, although this practice is not universally followed.

The next step in making a frequency table is to read the

scores, or have them read aloud slowly by an assistant, and as each score is read to make a mark or check of some sort opposite that size of score in the column previously prepared. A few persons prefer to make a cross (×), a period (.), or a check (✓) to record each occurrence of a score of a given size, but the most common scheme is to make a straight perpendicular mark, like an Arabic numeral 1, for each score as it is read. As soon as a given score is read for the fifth time, a diagonal tally mark is drawn through the four previous marks, tying them together in a group representing five occurrences of that particular score (𝍸). Tallying by fives makes the final counting much simpler.

The following ages "at last birthday" were reported by the pupils in a sixth-grade class and may well serve as illustrative material: 14, 15, 13, 11, 12, 10, 11, 12, 12, 12, 13, 14, 13, 11, 12, 12, 11, 11, 12, 12, 13, 12, 11, 14, 16, 13, 12, 12, 14, 13, 15, 12, 14, 11, 12, 10, 13, 13.

Under *A* in Table X is shown the arrangement of the column listing the different ages from high to low. Column *B* shows the result of making the check and tally marks representing the ages of the thirty-eight pupils, and *C* shows the summary in Arabic numerals of the marks appearing under *B*. In reporting a tabulation of scores it is not usual to show the separate check marks appearing under *B*, for the summary of frequencies under *C* requires less space on a page and furnishes the important numerical facts.

There are occasionally good reasons for using points or crosses in place of the perpendicular lines in tabulating frequencies. A table of frequency made carefully with

TABLE X

TALLY METHOD OF DISTRIBUTING AGES OF SIXTH-GRADE PUPILS

A AGE[1]	*B* FREQUENCY	*C* SUMMARY
16	|	1
15	||	2
14	𝍸	5
13	𝍸 |||	8
12	𝍸 𝍸 |||	13
11	𝍸 ||	7
10	||	2
Total	38	38

crosses of the same size has certain graphic advantages over those made with the straight lines and tally marks, but the results are a little more difficult to count or summarize. Table XI is, for example, almost as graphic in its presentation of the general age characteristics of this sixth-grade

TABLE XI

GRAPHIC TABULATION OF AGES OF SIXTH-GRADE PUPILS

AGE	FREQUENCY
16	×
15	××
14	×××××
13	××××××××
12	××××××××××××
11	×××××××
10	××

[1] Ages as at last birthday.

group as is Figure 13; but if it were desired to know how many pupils are twelve years old, it would be more difficult to count them from Table XI than from Table X.[1]

A semi-graphic tabulation of frequencies such as that in Table XI is not frequently used. Its greatest usefulness is found when it is desired to gain a general impression of a group of measures and where there is little probability that the facts will be needed for further calculations. It is usually more important to secure an accurate table of frequency from which a frequency surface or detailed

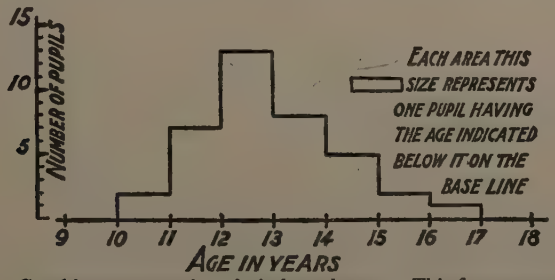


FIG. 13. — Graphic representation of sixth-grade ages. This figure represents the distribution of age scores given in Table X and Table XI.

statistical calculations may later be made than it is to tabulate the facts graphically from the beginning.

When the number of age scores or other records to be tabulated is too large for convenient classification by the methods described above, the information may be punched on cards and sorted by machines built especially for the purpose. The Hollerith Sorting Machine and the Powers Sorting Machine are most commonly used. The cards used by a sorting machine are of an exactly uniform size,

[1] Figures 6, 7, and 8 in Chapter II are excellent examples of graphic tabulations.

and are very carefully printed to show many columns of numbers, each column ranging from 0 down to 9, as in the schematic representation shown below.

SCHEMATIC REPRESENTATION OF CARD USED IN A SORTING MACHINE

GENERAL FACTS REGARDING PUPIL										READING				ARITHMETIC				WRITING			
Age		Class			School		Attendance														
										Speed		Power		Speed		Power		Speed		Quality	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Each column and each digit is in a certain specified location on the card and is assigned a definite significance. If the first two columns are to represent age, for example, each twelve-year-old child's card is punched at 1 in the first column and at 2 in the second; and each fifteen-year-old child's card is punched at 1 in the first and at 5 in the second column. The scores of the pupils in attendance, in deportment, or in any educational test may be punched in assigned columns on the same cards in a manner similar to that used in recording their ages.

The holes punched in a given card allow electric contacts to be made when the card in passing through the machine arrives at the proper place. By the closing of the electric circuit the card is automatically thrown out of the

general path into a pocket where all other cards similarly punched have fallen. When the entire lot of cards has been passed through the sorting machine, all those having a given score will be found in the same pocket. By counting the cards in each pocket one is able to fill in the items on the frequency table, just as one would from counting the number of check marks opposite each score if the tabulation had been done by hand.

Sorting machines work with great speed and accuracy, but the expense of renting the machines and of purchasing and punching cards make it inadvisable to use this method of tabulation unless there is a large number of different classifications to be made of the same large group of individuals. Comprehensive investigations of thousands of individuals should usually employ machines for classifying data, but studies involving only two or three facts about a few hundred individuals may ordinarily be carried on more economically by hand methods.

Percentile comparisons of groups of scores. — The comparison of one group of measures with another group should always include a tabulation of the number of cases by which each group is represented at each step or interval on the scale. The facts thus presented should usually be interpreted still further by graphic or other statistical methods. Table IX, for example, contains many significant facts regarding the differences in age between pupils of the fourth year at high school in two different cities, but without further treatment of the data it would be difficult to make interesting and helpful comments regarding these differences.

One useful type of comparison involves the statement of just what percentage of each distribution is below each

step on the scale. Table XII shows, for example, that .7 of one per cent of the pupils in City A are less than 16.0 years of age, while 2.4 per cent of the pupils in City B are below this age; that 12.5 per cent in City A and 18.3

TABLE XII

A Cumulative Frequency Table

Age at Last Birthday	City A			City B		
	No. of Pupils	Cumulative		No. of Pupils	Cumulative	
		No.	%		No.	%
23	1	304	100.	—	—	—
22	2	303	99.7	—	—	—
21	3	301	99.0	2	376	100.
20	22	298	98.0	4	374	99.5
19	46	276	90.8	26	370	98.4
18	113	230	75.7	111	344	91.5
17	79	117	38.5	164	233	62.0
16	36	38	12.5	60	69	18.3
15	2	2	.7	9	9	2.4
14	—	—	0	—	—	0

per cent in City B are below 17.0 years; and that 38.5 per cent in A and 62.0 per cent in B are below 18.0 years of age. Such comparisons showing what percentage of each distribution is below each step or interval on the measuring scale are very helpful and are fairly easily made.

The student will observe that one column in Table XII gives the frequency of each score, just as it was given in Table IX. The next column, however, gives for each step the sum of the frequencies on that step and all lower steps of the distribution. This column may be called the "cumulative frequency column," for it reports the cumulative

sum of all cases in that interval and lower intervals. It is obtained by adding up the cases in the ordinary frequency column, beginning at the bottom of the column. The numbers in the next column to the right show what percentages of the total group are represented by the figures in the cumulative frequency column. In City A, for example, there is one pupil at age 23, which makes a total of 304 pupils below 24.0 years, but 304 is 100.0% of the total number of pupils in the distribution. Similarly, on the next lower line 2 pupils are called "twenty-two," which makes a total of 303 pupils (or 99.7% of all) who are below 23.0 years of age. The digits in the percentage column are obtained by dividing the digits in the cumulative frequency column by the total number of pupils in the distribution.

From the cumulative percentage column for City A one may state that 24.0 on the age scale is the 100 percentile point in the distribution for this school, that 23.0 years is the 99.7 percentile point, and that 22.0 years is the 99.0 percentile point. The statement that "21.0 years of age is the 98.0 percentile point for City A" means that exactly 98.0 per cent of the total group is lower than this point on the age scale. In other words, the percentile value in a distribution held by any given point on the measuring scale indicates just what percentage of the entire distribution is below that point on the scale. In Figure 14, therefore, exactly 98.0 per cent of the area of the frequency surface is at the left of the dotted line erected at 21.0 years on the scale at the base of the figure. In a similar manner, 12.5 per cent of the total area is below age 17.0 years.

The comparison of two or more distributions of scores

on the same scale is greatly facilitated by this simple determination of what percentage of each distribution is below each class or interval on the original measuring scale. It is not difficult to understand the significance of the statement that 18.0 years on the age scale is the 38.5 percentile point for City A but the 62.0 percentile point for

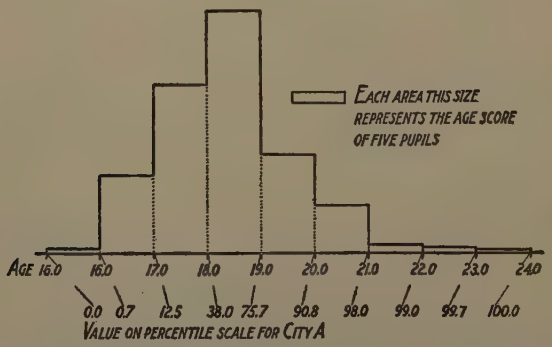


FIG. 14. — Percentile values of year limits on measuring scale for age. This figure presents exactly the same facts for City A as were given in Tables IX and XII.

City B. One obtains information that is really useful when told that 19.0 years on the age scale has a 75.7 percentile value in City A and a 91.5 percentile value in City B.

The comparisons made possible by calculations of the type presented in Table XII may not fully satisfy one's needs. It may be desirable to select certain important percentile scale points and determine what points on the measuring scale represent them in each different distribution. In other words, one may wish to know what points on the age scale represent the 90.0 percentile point in the two distributions. It may be important to learn at just what point between 18.0 and 19.0 years one would find

the 50 percentile point for the City A distribution. The cumulative frequency table and its percentages show what proportion of a group is below any given step or interval on the measuring scale, but it is frequently even more desirable to discover exactly what point on the measuring scale has below it a given proportion of the total distribution.

The calculation of the points on the measuring scale which correspond to given percentile points in a distribution of scores is especially important when one desires to describe the extent to which a group of scores is scattered or massed. The statement that ages in City A range from 15 to 23 inclusive while in City B they range from 15 to 21 inclusive is not an adequate description of the greater compactness and uniformity of the senior class ages in City B. A few extremely old or a few young pupils would make radical changes in the extreme limits of a distribution even when the majority of the ages were included in a very narrow range.

A favorite method of describing how widely a distribution is dispersed or scattered consists in cutting off a certain proportion of the extreme cases at either end and then stating the limits within which the remainder of the distribution is found. By cutting off the highest fourth of the distribution and the lowest fourth, for example, one may state the limits of the " middle fifty per cent " of the distribution and compare these limits with the corresponding limits in another distribution. The 25 percentile point in the distribution for City A is 17 years, 5 months, and 23 days, while the 75 percentile point is 18 years, 11 months, and 20 days. Between these two limits one would find the middle fifty per cent of the ages in this dis-

tribution. In the distribution for City B, the middle fifty per cent would be found between 17 years, 1 month, and 25 days (the 25 percentile point) and 18 years, 5 months, and 9 days (the 75 percentile point). The middle half of the ages in City A therefore spreads out over a range of 1 year, 5 months, and 27 days, while the middle half in City B covers only 1 year, 3 months, and 14 days.

Occasionally an investigator wishes to report the range of the " middle 80 per cent " or of the " middle 90 per cent " in a distribution. By cutting off the lowest ten per cent and the highest ten per cent one may eliminate the extreme cases and yet leave a large majority of the distribution. The 10 percentile in City A, for example, is at 16 years, 9 months, and 14 days, while the 90 percentile is at 19 years, 11 months, and 11 days. The range of the middle 80 per cent therefore covers a period of 3 years, 1 month, and 27 days in City A. The corresponding range in City B covers only 2 years and 7 months. Although the wider dispersion of ages in City A is more evident from a comparison of the ranges of the middle 80 per cent than from a comparison of the middle 50 per cent ranges, there are statistical and practical reasons for measuring dispersion by the middle 50 per cent range more frequently than by any other per cent range.

The percentile points in a distribution divide it into a hundred equal parts. The 10 percentile point is also known as the first decile point, the 20 percentile point as the second decile point, and so on. The decile points divide the distribution into ten equal parts, just as the percentile points divide it into one hundred equal parts. The sixth decile point therefore has six-tenths of the total distribution

below it, and the ninth decile point has nine-tenths of the distribution below it.

The 25 percentile point, the 50 percentile point, and the 75 percentile point are also known as the first quartile point, the second quartile point, and the third quartile point, respectively. The 25 percentile point is called the first quartile point (Q_1) because it has just one-fourth or quarter of the distribution below it. The 50 percentile point has two quarters or one-half of the distribution below it and is therefore known as the second quartile point (Q_2), or median point. The 75 percentile point has three quarters of the distribution below it and is known as the third quartile point (Q_3). The quartile points divide the distribution into four equal quarters, just as the percentile points divide it into a hundred equal hundredths, and the decile points divide it into ten equal tenths. The following scheme may help the student to keep in mind the relationships of these various points in a distribution.

100	percentile point	=	10th decile point	
90	" "	=	9th " "	
80	" "	=	8th " "	
75	" "	=	———	3rd quartile point
70	" "	=	7th " "	
60	" "	=	6th " "	
50	" "	=	5th " "	= 2nd " "
40	" "	=	4th " "	
30	" "	=	3rd " "	
25	" "	=	———	1st " "
20	" "	=	2nd " "	
10	" "	=	1st " "	
0	" "	=	0 " "	

Figure 15 shows the same surface of frequency for the ages of high school seniors in City A as was shown in Figure 14, but in Figure 15 the dotted perpendicular lines are erected at the decile points for the distribution rather than at the full year points on the age scale. The surface of frequency in Figure 15 is divided into ten equal parts by the dotted lines erected at the decile points on the base of the figure. Dashed lines have also been erected at the

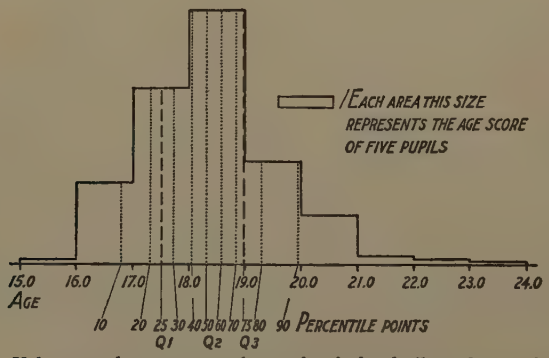


FIG. 15. — Values on the age measuring scale of the decile and quartile points in the distribution of high school seniors' ages in City A.

first and third quartile points on the base of Figure 15 to show how a distribution is divided by the quartile points into four equal quarters.

The differences between the decile points in the distribution for City A and the corresponding decile points for City B are represented in Figure 16. At the 20 percentile values City A pupils are only three months older than pupils in City B, but at the median or 50 percentile point pupils in City A are seven months older than pupils in B, and at the 90 percentile point pupils in City A are

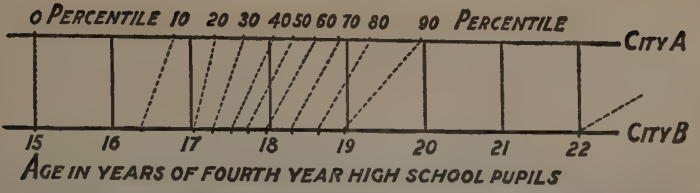

0 PERCENTILE 10 20 30 40 50 60 70 80 90 PERCENTILE
CITY A
CITY B
15 16 17 18 19 20 21 22
AGE IN YEARS OF FOURTH YEAR HIGH SCHOOL PUPILS

FIG. 16. — Comparison by deciles of fourth-year high school ages in two cities. Note that the two cities are more nearly alike in their lower deciles than in their higher.

almost a year older than those in B. The same facts are perhaps more adequately represented in Figure 17, in which the decile points are plotted along one axis while the age facts are plotted along the other axis of the figure. It would have been possible to have plotted certain additional points on each curve from the percentile values obtained in Table XII. Such points would have been intermediate between various decile points, but would not have altered the general shapes of the curves, except perhaps near their ends. As Figure 17 stands, one may either read from it the age score corre-

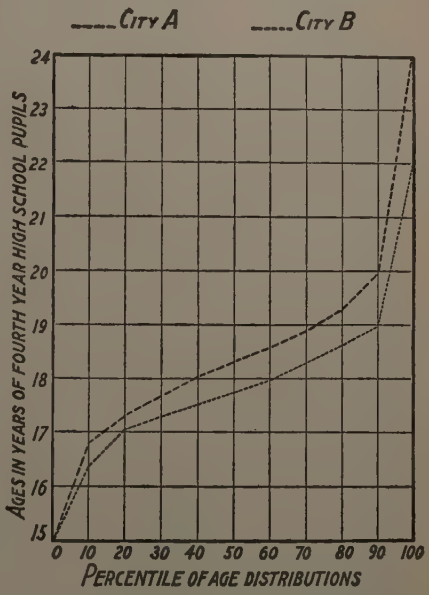

CITY A
CITY B
AGES IN YEARS OF FOURTH YEAR HIGH SCHOOL PUPILS
24 23 22 21 20 19 18 17 16 15
0 10 20 30 40 50 60 70 80 90 100
PERCENTILE OF AGE DISTRIBUTIONS

FIG. 17. — Decile curves for ages of high school seniors in cities A and B. Note that in each city the curves are almost straight between the 20 percentile and the 80 percentile.

sponding to any decile point of either distribution, or one may estimate from it with ease the percentile values of any given age.

The calculation of percentile points from distributions of scores. — One of the most fundamental operations in all statistical work is the calculation of percentile points from frequency tables. Fortunately, one percentile value is calculated by exactly the same steps and on exactly the same assumptions as another. One who has learned, therefore, to calculate a 10 percentile point, a 50 percentile point, or a 25 percentile point, does not need to learn anything new in order to calculate a 75 percentile point, a 2 percentile point, or a $12\frac{1}{2}$ percentile point. It is important, however, that any one who plans to employ standard classroom tests should understand thoroughly and should have considerable skill in the calculation of percentile points from tables giving distributions of scores.

The most satisfactory method of learning to compute a percentile point is to go through the actual process several times. One may begin by calculating the 20 percentile point in the distribution of ages for high-school seniors in City B. The 20 percentile point is, by definition, a point on the age scale so located that 20% (or one-fifth) of all the 376 pupils are younger, while 80% (or four-fifths) of them are older. Since the actual age of each individual pupil is not given, and since it is impossible to divide 376 exactly by 5, it is evident that one must by calculation make the most accurate estimate possible of a point which would cut off the lowest 20% of the 376 pupils. Certain things will have to be assumed with regard to the ages of some of these pupils, and as careful an estimate

as possible will have to be made of the point on the age scale below which there are 75.2 pupils and above which there are 300.8 pupils.

To save the reader from turning back repeatedly to Table IX to observe the distribution of ages in the fourth-year high-school class of City B, it is reproduced here as Table XIII. In order to exceed the age of the lowest 75.2 of these pupils, one must first pass over the 9 pupils who are 15 years old and the 60 who are 16 years old. Having passed the lowest 69 pupils in this manner, one is ready to pass the next lowest 6.2 pupils (75.2 − 69 = 6.2), who have registered with the total of 164 pupils at age 17.

TABLE XIII

AGES OF FOURTH-YEAR HIGH-SCHOOL PUPILS

AGE AT LAST BIRTHDAY	NO. OF PUPILS CITY B
21	2
20	4
19	26
18	111
17	164
16	60
15	9
All ages	376

What proportion of the total range of ages registered as 17 years old is included by taking 6.2 pupils from the total of 164 pupils who are of this age? The answer to this question is fundamental in all statistical work. Where it is not accurately known how a certain number of cases are distributed over a given range in a scale, the assumption may be made that no two cases are exactly equal, but that the range is divided equally among the cases appearing on it.[1] If there are two pupils 17 years of age at their last birthday, one may be assumed to be between 17 years, 0 months, 0

[1] This assumption was more fully discussed in the preceding chapter in connection with the calculation of mean or average age, and in the section on the "continuity of time." (See pp. 90–92.)

days and 17 years, 6 months, 0 days; while the other may be assumed to be between 17 years, 6 months, 0 days and 18 years, 0 months, 0 days. If there are three pupils 17 years old at their last birthday, the youngest may be thought of as being from 17 to $17\frac{1}{3}$, the next from $17\frac{1}{3}$ to $17\frac{2}{3}$, and the oldest from $17\frac{2}{3}$ to 18 years old.

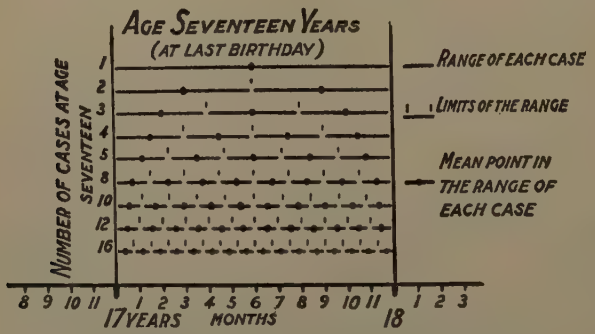


FIG. 18. — Graphic representation of the assumption that all cases on a given step are distributed evenly over the total range of the step.

These and other assumptions of a similar nature are represented in Figure 18.

Since there are 164 pupils registered in City B as 17 years old, the youngest should be thought of as being between 17 and $17\frac{1}{164}$ years old, the next youngest as between $17\frac{1}{164}$ and $17\frac{2}{164}$, and so on. If 6.2 of the youngest are passed, therefore, one will have arrived at a point on the age scale represented by $17 + \frac{6.2}{164}$ years; which is 17.037 years, or 17 years, 0 months, 13.6 days. The 20 percentile point in City B's distribution of ages for the fourth-year high-school pupils is thus found to be 17 years, 0 months, and 14 days. This age is the safest

estimate one can make of the point on the age scale below which one would find the ages of one-fifth of the pupils and above which one would find four-fifths of them.

As another example, let us suppose that one wishes to know the 95 percentile of the distribution of ages in Table XIII. The 95 percentile is a sort of practical upper limit to any distribution, in the sense that there is only one chance in twenty that a given child selected at random from the group would be so high in the scale. To pass the lowest 95% of 376 cases would mean passing 357.2 of them. By the time one has passed the 9 scored at 15 years, the 60 scored at 16 years, the 164 scored at 17 years, and the 111 scored at 18 years, he has passed 344 cases and needs to take the lowest 13.2 of the 26 cases scored at 19 years. Assuming that the 26 cases are each allowed an equal portion of the range from 19 years, 0 days, up to 20 years, 0 days, the lowest 13.2 cases will cover $\frac{13.2}{26}$ of the total year, and $\frac{13.2}{26}$ year is .508 of a year, or 6 months and 2.7 days. The point being sought (the 95 percentile) is therefore 19 years, 6 months, and 3 days, — that is, 6 months, 3 days older than the youngest possible 19-year-old pupil.

SUMMARY OF CALCULATION

Total no. cases,	376
1% of cases,	3.76
95% of cases,	357.2
Total below 19,	344.0
To be taken at 19,	13.2
Total at 19,	26

$\frac{13.2}{26} = .5077$ of the year

$19.0 + .5077 = 19.508 =$ 19 years, 6 months, 3 days.

Calculation of the median point. — The most popular measure of the average or central tendency in a group of

scores on an educational test is the median point, which is nothing more than the 50 percentile point and is calculated in exactly the same manner as any other percentile point. The special advantages of the median point will be discussed in a later chapter, but it may be readily seen that, since in any distribution there are just as many scores smaller as there are scores larger than the 50 percentile point, it stands very well as a single representative of any group. One additional illustration of the calculation of a percentile point may therefore serve at the same time to illustrate the calculation of a median point.

The median or 50 percentile point divides a distribution into two exactly equal parts. To divide a group of 376 pupils into two equal parts, one would need to pass 188 of them. When one has passed the 9 who are 15 years old and the 60 who are 16 years old, one will still need to pass 119 (188 − 69 = 119) of the 164 who are 17 years old in order to have passed a total of 188. To pass 119 of the 164 means to cover $\frac{119}{164}$ of the range called age 17, and $\frac{119}{164}$ of one year would be .726 of a year. Since 17 years old means from 17.0 up to 18.0, the median or 50 percentile age of pupils in this distribution is 17.726 years, or 17 years, 8 months, and 21 days.

SUMMARY OF CALCULATION

Total no. cases, 376
1% of cases, 3.76
50% of cases, 188
No. below 17 years, 69
To be taken at 17, 119
No. at 17 years, 164
$\frac{119}{164} = .726$
17 years = 17.0 to 18.0
17.726 = 17 years, 8 months, 21 days.

Rules for finding a percentile point may prove more

confusing than helpful to some of those who try to use them. The three foregoing illustrations should be studied carefully by those who wish to learn to make the calculations. One may then read and perhaps understand the following summary of the different steps in the process:

(1) Determine the number of individual scores or frequencies to be passed.

(2) Count into the distribution to find the distribution class or step in which the desired point lies, and to determine how many of the frequencies in that step must be passed.

(3) Calculate the fraction of the total step represented by the frequencies that are to be passed.

(4) Add this fraction to the lowest value of the step. (In cases where one step on the distribution is not exactly one unit on the scale of measurement, it will be necessary to convert both the fraction found in 3 and the lowest value of the step into terms of the measuring scale units before making the final addition.)

Confusion has sometimes arisen in the minds of persons learning to calculate percentile points from a distribution of scores, especially in the calculation of the median or 50 percentile point. The chief source of such confusion has been the fact that certain writers on statistics have intended to indicate only the median *score* when they have spoken of " the median," and to indicate the 25 percentile *score* when they have spoken of " the 25 percentile." The median score is, of course, the score in which the median point is found, just as the 75 percentile score is the score in which the 75 percentile point is located. The present writer has found it more satisfactory to teach students to locate percentile points only. The introduction of directions and formulæ for the calculation of percentile steps or scores usually causes confusion and is not necessary.

If the reader will master the process of locating percentile points, he will have no difficulty whatever in finding the step or group of scores in which such a point is located.

Summary. — Accurate measurement of the individuals composing a group is only the first step in the comparison of the group with other groups measured by the same scale. The mean, obtained by adding all the individual measurements together and dividing by their number, is a convenient single index, but is not fully representative of an entire group and is frequently misleading as to the character of the group as a whole. A distribution of the measurements of a group, showing how many individuals made each different score, is the most satisfactory basis for comparing one group with another. Graphic representations of such distributions are effective in giving an easily comprehended picture of the gross differences between groups. For making exact statistical statements of differences, the calculation of various percentile points is desirable. The 50 percentile or median point is a particularly convenient point of reference in any distribution of scores.

QUESTIONS AND EXERCISES

1. State as well as you can why the mean of a group of measurements is not always adequate as a representative of the entire group.

2. Prepare a frequency table to show the facts regarding the ages of pupils in a class composed of 20 pupils whose ages, as of last birthday, are as follows: 10, 13, 11, 15, 13, 14, 12, 12, 11, 13, 14, 15, 12, 14, 13, 16, 14, 12, 13, 14.

3. What is the mean age in years, months, and days of pupils in the above class?

4. Prepare a surface of frequency to represent the distribution for the above class.

5. Calculate the 25 percentile, median, and 75 percentile points in the above distribution, and divide the surface of frequency (prepared in 4) into its four quarters.

6. What are the limits of the middle 80 per cent of the 20 pupils? Indicate these on the base line of the surface of frequency.

7. Calculate the 5 percentile point of the 304 ages tabulated in Table IX for City A. What would this 5 percentile be if the ages in Table IX had been to the nearest birthday rather than to the last birthday?

8. What is the median point in the ages of the sixth-grade pupils listed in Table X? Indicate this point in Figure 13.

REFERENCES FOR READING

W. P. and E. M. Elderton, *Primer of Statistics*, Chapters I–III.
T. L. Kelley, *Statistical Method*, pp. 54–59.
Strayer and Norsworthy, *How to Teach*, pp. 277–284.
H. O. Rugg, *Statistical Methods Applied to Education*, pp. 103–114.

CHAPTER VI

A TEST FOR SPELLING ABILITY

SCHOOL pupils have probably been tested more frequently for ability in spelling than for ability in any other subject. The most common method of reciting spelling lessons is by giving a written or oral spelling test. One of the most popular types of intellectual competition between schools, especially in the rural sections of America, has been the "spelling contest." Reasons for the widespread popularity of tests in spelling may perhaps be found in the brevity of the units by which ability in spelling is measured, the ease with which accuracy may be checked, and the apparent certainty with which relative ability in this field may be determined.

Words spelled are not equally important as measures of spelling ability. — In the old time "country spelling bee" where one side of the room "spelled down" the other, the social interests and relations were more important than the intellectual, but even the intellectual elements were decidedly unequal in their importance. The side which won was not necessarily the side which had the highest average in spelling ability but rather the side on which the *best* spellers were better than the best the other side could produce. Victory was decided not by the relative averages of the two groups but by the relative abilities of the upper five per cent of spellers in each of the

two groups. Similarly, words of average difficulty had little to do with the final result of such a contest, for the final decision rested with the side whose best spellers were able to spell harder words than could be spelled by the best spellers among their opponents. The most " deadly " words in such a spelling warfare were those near the end of the spelling book.

In the ordinary written test, by means of which teachers try to discover whether their pupils have mastered a spelling lesson, each word is given equal weight. If the list contains ten words, 10% is deducted from 100 for each word incorrectly spelled; and if the list contains twenty-five words, 4% is " taken off " for each error. The first attempts to compare schools with each other by means of a uniform spelling test followed the same arbitrary rule of allowing an equal amount of credit for each properly spelled word. In the investigation conducted by Dr. J. M. Rice,[1] mentioned in Chapter III, this plan of considering all errors equally important was used.

Among the words used by Dr. Rice were the following: *changeable*, *necessary*, *certainly*, *whistling*, *purpose*, *neighbor*, *laughter*, *weather*, *because*, and *picture*. A mere glance at these words would convince any one who has taught spelling in the elementary grades that they are not equally difficult, and yet in the absence of experimental evidence Dr. Rice could not do otherwise than consider them of equal importance. Professor Thorndike a few years later reported the frequency with which each of the words used by Rice had been incorrectly spelled in an actual trial

[1] J. M. Rice, "The Futility of the Spelling Grind," *The Forum*, 23, 163–172, 409–419.

with boys and girls in a fifth grade.[1] Figure 19 shows the relative frequencies of errors reported by Thorndike on each of the ten words. It would, of course, be absurd to conclude from these results that *neighbor* is ten times as hard to spell as *because*, or half as hard as *whistling*, but the scores in Figure 19 certainly demonstrate that not all words are equally important as measures of spelling ability.

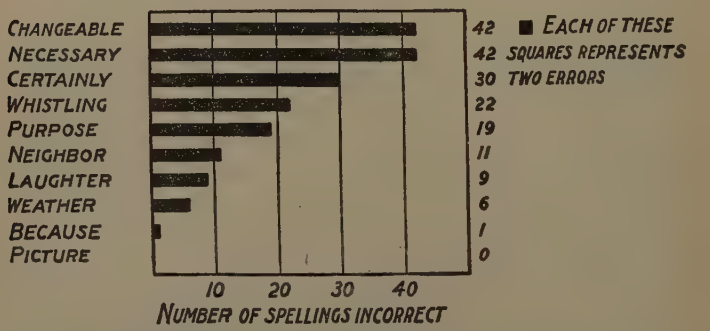


FIG. 19. — Unequal difficulty of words used in same spelling test as of equal difficulty. The results given above are for a 5A grade tested by Thorndike. The ten words are chosen from the fifty-word list used by Dr. Rice.

Ten, twenty, thirty, or forty times as many pupils may fail on one word as on another in the same spelling test.

In commenting on the spelling list used in the Springfield Tests (page 67), it was suggested that few modern teachers would care to have their efficiency as teachers measured by the success of their pupils in spelling the words of that list. Such words as *chirography*, *evanescent*, and *feignedly* are seldom found in the written or spoken vocabularies of school pupils. These words should there-

[1] E. L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements*, p. 8.

fore be considered less important as measures of spelling ability than the words which these pupils are constantly speaking and writing. The importance of teaching children to spell a given word, or of including it in a scale for the measurement of the results of spelling instruction, is proportional not only to the inherent difficulty of the word itself but also to the frequency with which the children will be called on to spell the word in their daily work. Whether measured by the frequency of their use or by their natural spelling difficulties, words differ tremendously in their importance as indices of spelling ability.

The conditions of a spelling test affect the results. — When the information is received that one-seventh of the pupils in a certain fifth grade were unable to spell the word *peace*, one is inclined to think he knows something definite regarding the spelling ability of that fifth grade. One would be tempted, for example, to assume that another fifth grade, of which one-fourth is reported as unable to spell *peace*, was less successful in spelling this word. Such a conclusion would not necessarily be correct, however, unless the conditions of the test were known to be identical. " P-i-e-c-e " may have been accepted as a satisfactory spelling in the case of the first group and not accepted with the second group. And even though " p-e-a-c-e " was the only accepted spelling in both cases, the teacher may have pronounced the word so poorly in the second class that many of the pupils thought she said " peas." The teacher may have pronounced the word twice or may have explained its meaning to the first class, while the second class may have had only one opportunity to hear the word pronounced and little time in which to write it. In one

of the classes the word may have been dictated in a sentence with other words to be written, while in the other it may have been pronounced by itself for oral spelling. Every difference in test procedure or conditions may affect the results materially.

Dr. Rice and Superintendent Cornman, whose spelling studies have previously been mentioned (pp. 64 and 71), both used what are known as Composition Spelling Tests, in which the pupils were asked to write compositions on assigned topics and were then rated according to the accuracy of their spelling in these compositions. In such tests and in Cornman's other test, where the pupils were asked to write "as many words as you can" in 15 minutes, the percentage of spelling accuracy increased regularly from about 93% in the third grade to about 98% in the eighth grade, but this improvement in percentage was perhaps due as much to better judgment in the upper grades as to superior spelling ability. No two pupils used the same words, and it is probable that each used the words which were most familiar to him. The pupil of poor spelling ability but good judgment might in such a test write a hundred familiar words correctly, while the pupil of higher spelling ability but less caution might write two hundred words, including many that were incorrectly spelled, thereby obtaining an inferior score.

In order to make valid comparisons on the basis of test results, the same scale must be used, and the conditions must be the same. In spelling, either the same list of words must be used, or an equivalent list in which each word is of exactly the same difficulty as the corresponding word in the original list. The methods by which pupils are made

aware of the words to be spelled must be the same for all, and the time allowed for writing each word must be clearly defined and carefully observed. Each variation in spelling that is to be considered correct must be specified. In so far as the conditions under which one group of pupils is tested are more favorable than the conditions for another group, comparison of spelling abilities on the basis of these results is open to question.

With these considerations in mind, the writer has prepared and used a test in spelling which makes use of words selected from the Buckingham Extension of the Ayres Spelling Scale.[1] The test directions, spelling lists, and standard scores will be presented before the special character of the lists is discussed.

Directions for a standard classroom test in spelling. — This test should be given by a teacher or supervisor with whose voice, manner, and pronunciation the pupils are familiar. Although the test is designed especially for the measurement of pupils in Grades III to VIII, it may be used in Grade II or in Grade IX if desired. The test is intended to measure the spelling power of the pupils rather than their familiarity with the spelling of words that are appropriate only to their particular level of ability. The entire list of words should therefore be used in each test, even though certain words in the list seem entirely inappropriate.

[1] The Buckingham Extension of the Ayres Spelling Scale contains 505 words that were not in the original Ayres Spelling Scale, the majority of these additional words being in the upper or more difficult ranges of the scale. The original Ayres Scale may be procured from the Russell Sage Foundation, 130 East 22nd Street, New York City. The Buckingham Extension may be obtained from the Public School Publishing Co., Bloomington, Ill.

See that all pupils in the room have convenient tables or desks on which to write and that all are supplied with pencils and with paper of a uniform size.

Have each child write his name in the upper right-hand corner of his paper. (Demonstrate the location on a sheet of the uniform paper.) As soon as the names have been recorded, say to the pupils: "You are in the 4A grade (or in the 6B grade, as the case may be). Write '4A Grade' just below your name; and then, just below the word 'Grade', write your age at your last birthday. How old were you the last time you had a birthday? See that you have your name, your grade and your age in the upper right-hand corner of the sheet."

When this identifying information has been recorded, make the following statement: "This will be a spelling test. There are twenty-five words in the list to be spelled, so you may make the numbers from one to twenty-five in a column down the left-hand side of the page. (Demonstrate the location of this column on a sheet of the spelling paper.) Write the numbers quickly: from *one* down to *twenty-five*." Allow forty or fifty seconds for writing these numbers.

When the numbers are written, the examiner should explain the test as follows: "The first words in the list are very easily spelled, but each word is more difficult than the one before it. The last words are perhaps too hard for most of you, but you must do your best on each word.

"I shall speak each word twice, but never more than two times. Watch me as I pronounce the word, and do not begin writing until I have pronounced it the second time. Be sure to write each word after its proper number in the column you have prepared. Listen closely until I have pronounced each word the second time. First word: 'yes, yes.'"

Allow ten seconds for writing the word, and then call the next word. "Second word: 'face, face.'" And so on for the entire list of twenty-five words.

Allow approximately one and one-half seconds between the first and second pronunciation of each word. For spelling each of the first fifteen words, allow from ten to twelve seconds; and for spelling each of the last ten words, allow approximately fifteen seconds. As

soon as the entire list of twenty-five words has been pronounced and written, collect the papers and make any special notes that seem necessary regarding the test conditions. Such special records should be kept tied or otherwise connected with the package of test papers from the room where the conditions were discovered.

The standard spelling test lists. — The first word in each list is taken from column H of Buckingham's Extension

STANDARD SPELLING TEST LISTS

(Buckingham-Ayres, H–AF)

COLUMN	LIST 1	LIST 2	LIST 3	LIST 4
H	1. yes	1. box	1. baby	1. day
I	2. face	2. tree	2. paper	2. foot
J	3. forget	3. girl	3. train	3. sister
K	4. water	4. clean	4. horse	4. party
L	5. wish	5. world	5. fight	5. table
M	6. railroad	6. contract	6. children	6. teach
N	7. yesterday	7. bridge	7. number	7. chain
O	8. liberty	8. company	8. public	8. second
P	9. entrance	9. beautiful	9. fortune	9. rapid
Q	10. property	10. relative	10. famous	10. publish
R	11. visitor	11. prepare	11. service	11. salary
S	12. examination	12. investigate	12. supply	12. local
T	13. political	13. conference	13. accident	13. recent
U	14. receive	14. agreement	14. material	14. citizen
V	15. secretary	15. experience	15. reference	15. campaign
W	16. athletic	16. practical	16. character	16. science
X	17. committee	17. annual	17. artificial	17. peculiar
Y	18. restaurant	18. orchestra	18. accuracy	18. decision
Z	19. porcelain	19. enthusiasm	19. judgment	19. allege
AA	20. immediately	20. apologize	20. municipal	20. menagerie
AB	21. guarantee	21. pneumonia	21. villain	21. lieutenant
AC	22. cantaloupe	22. syndicate	22. sovereign	22. embarrassment
AD	23. souvenir	23. chauffeur	23. penitentiary	23. maneuver
AE	24. rhinoceros	24. inflammable	24. hallelujahs	24. rhinoceros
AF	25. conscientious	25. dissension	25. rendezvous	25. naphtha

STANDARD SPELLING TEST LISTS

(Buckingham-Ayres, H–AF)

COLUMN	LIST 5	LIST 6	LIST 7	LIST 8
H	1. well	1. door	1. soft	1. five
I	2. winter	2. river	2. plant	2. white
J	3. seven	3. fire	3. room	3. town
K	4. first	4. lady	4. stamp	4. glass
L	5. soap	5. black	5. church	5. summer
M	6. please	6. dress	6. build	6. member
N	7. dollar	7. history	7. doctor	7. always
O	8. judge	8. request	8. appear	8. perfect
P	9. desire	9. enjoy	9. human	9. several
Q	10. position	10. surprise	10. private	10. declare
R	11. avenue	11. promise	11. feature	11. secure
S	12. century	12. motion	12. serious	12. often
T	13. circular	13. absence	13. victim	13. various
U	14. divide	14. relief	14. difficulty	14. earliest
V	15. mysterious	15. evidence	15. testimony	15. license
W	16. calendar	16. separate	16. proceed	16. extreme
X	17. familiar	17. convenient	17. immediate	17. opportunity
Y	18. possess	18. surgeon	18. rèsem-blance	18. immense
Z	19. physician	19. parallel	19. millinery	19. recommend
AA	20. privilege	20. discipline	20. appropriate	20. chandelier
AB	21. combustible	21. occurrence	21. incessant	21. proficiency
AC	22. sovereign	22. poultice	22. syndicate	22. embarrass-ment
AD	23. appendicitis	23. hippopota-mus	23. miscella-neous	23. souvenir
AE	24. inflam-mable	24. rhinoceros	24. hallelujahs	24. inflam-mable
AF	25. jardiniere	25. discernible	25. conscien-tious	25. rendezvous

of the Ayres Spelling Scale, the second word from column I, the third from column J, and so on, one word from each column, the last word in each list being from column AF, the last and most difficult column in the scale. Additional lists, equivalent to these in every respect, may be prepared by any one who cares to select them in the same manner from columns H to AF of the Buckingham-Ayres lists.

Directions for assigning the scores. — Each pupil's score in this test is the number of words correctly spelled. Any error that makes the spelling less than perfect should serve to remove the word in which it occurs from consideration as a correctly spelled word. A small error in the spelling of a word is considered just as serious as a large one. If the pupil's version is a correct spelling of the word according to any standard modern English dictionary, it should receive full credit. The omission of the dot over an *i* or of the cross stroke on a *t* should be called an error. If the pupil has made a correction, consider only the corrected spelling. When it is impossible to tell which is intended to be the final form, give credit for the correct spelling if either is clearly correct. Where it is not possible to read the pupil's handwriting, consider the spellings incorrect.

Since there are twenty-five words in each list, the maximum score obtainable in this test is 25. An individual pupil should not be judged as to his spelling ability except by the average of the results of several lists, but the median result on a single list may usually be taken as a satisfactory index of the spelling ability of a grade or group of pupils.

Standard scores in spelling. — Classrooms in more than eighty cities were tested by Dr. Ayres in developing his spelling scale, and the results he has reported make it possible to compare the achievement of any classroom with an average calculated for each grade in the eighty cities. Results are also available for the white children of the elementary schools of Baltimore, Maryland, Greensboro, N. C., and two small cities in New Jersey. Table XIV presents these scores by grades, as standards with

which other grade groups may be compared. Table XV presents the results by age groups in the cities in which these lists have been used. Figures 20 and 21 present

TABLE XIV

GRADE SCORES ON STANDARD SPELLING LIST OF TWENTY-FIVE WORDS

(Columns H to AF, Buckingham-Ayres Scale)

GRADE	II	III	IV	V	VI	VII	VIII	IX
Average (80 cities)	5.0	7.6	10.5	12.5	14.5	16.4	18.3	20.0
Greensboro, N. C.	3.1	6.5	—	11.6	14.2	17.5	19.6	—
Baltimore, Md.	—	—	9.7	12.6	15.2	17.1	18.5	—
Hackensack, N. J.	3.3	7.0	9.2	12.5	14.9	17.4	19.3	20.7
West Orange, N. J.	2.6	5.5	8.6	11.3	13.8	—	—	—

TABLE XV

AGE SCORES ON STANDARD SPELLING LIST OF TWENTY-FIVE WORDS

AGE	7	8	9	10	11	12	13	14	15	16
Baltimore (Gds IV–VIII)	—	11.1	10.6	11.7	13.4	14.7	16.0	16.9	16.6	17.3
Greensboro (Gds II–VIII)	—	5.6	6.9	11.2	11.8	15.5	17.8	17.3	—	—
Hackensack (Gds II–VIII)	7.2	7.5	9.8	12.4	15.2	16.2	17.8	18.3	18.8	19.0
West Orange (Gds II–VI)	3.2	4.6	6.2	8.7	11.2	11.5	12.5	—	—	—

respectively the same results. It is unfortunate that no results by age groups are available for the eighty cities originally tested by Ayres. In comparing with the scores in Table XV and Figure 21, one should remember to observe that only the over-age (and therefore probably less capable) pupils were included in the 15- and 16-year-old

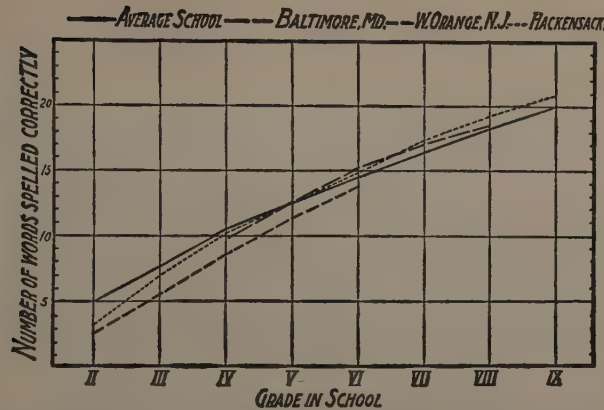


FIG. 20. — Comparative scores in spelling, by school grades. The "Average School" scores are calculated from the mean scores published on the Buckingham Extension of the Ayres Spelling Scale. The published scores on the Ayres Scale were obtained from tests in about 80 cities.

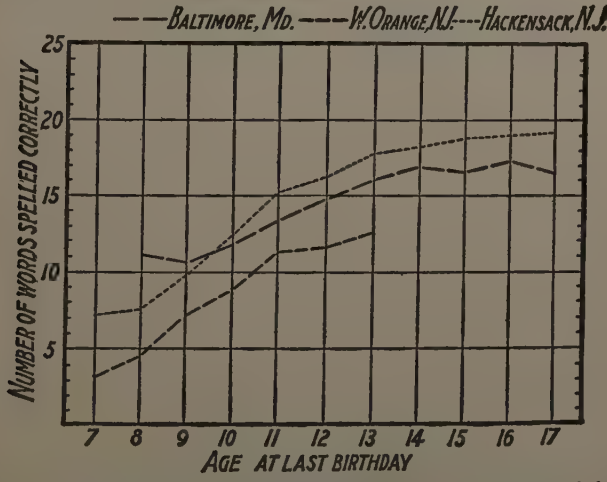


FIG. 21. — Comparative scores in spelling, by ages. It should be noted that only Grades IV to VIII were tested in Baltimore, and only Grades II to VI in West Orange. Only ages 10 to 14 are fair representations of Baltimore, therefore, and ages 8 to 12 of West Orange.

groups, and that only the under-age (and therefore probably more capable) pupils were included in Baltimore in the 8- and 9-year-old groups.

Special advantages of the standard spelling test lists. — The spelling words presented in the above lists differ in several ways from the words that have formerly been used for measuring spelling ability. They differ from the lists usually prepared by teachers and from the lists used by early investigators in that they are selected from the most commonly used words in the English language. The Ayres Spelling Scale is composed of the 1000 most frequently used words, as determined by Ayres and others in a careful study of personal and business letters, the Bible, and children's writing vocabularies. The 505 additional words contained in Buckingham's Extension of the Ayres Scale were chosen from those that are common to a large number of spelling books. The lists presented above were taken as far as possible from the words of the original Ayres Scale, in order that actual use in written work might be a more important determinant than frequent appearance in spelling books. In either case, however, the words of these lists may be taken as words which children should be able to spell.

The spelling test lists on pages 135 and 136 are not unique in being selected from the Ayres Spelling Scale. The Springfield (Ill.) Survey set the example by using words from this source in 1915, and practically every school survey since that time has used spelling words from the same source. In almost all cases, however, a different list has been used for each school grade, so that it was difficult to compare the spelling ability of an eighth-grade

pupil with that of a sixth-grade pupil, and absolutely impossible to secure a summary on an equal basis of the scores of all twelve-year-old pupils. The present lists were selected horizontally from the Ayres Scale, while the common practice has been to select them vertically. The present attempt is to measure spelling power, to find how difficult a word can be spelled, while the former effort was to find how many words of a given degree of difficulty could be spelled correctly. Each word in the present lists is just as much more difficult than the word before it as it is less difficult than the word after it.

The reader may make up as many additional lists as he wishes, by selecting twenty-five words in the same manner for each list. The comparative scores shown here will apply equally well to the new lists thus prepared. The possibility of many new spelling test lists of the same character and difficulty is in itself a great advantage, for it allows a verification of results through a repetition of the tests. It is furthermore of real importance that one be able to measure growth or improvement, — to compare the ability of a class or group now with its ability five years hence. This is easily possible with the increasing-difficulty scales here provided, but is not possible with lists of equally difficult words.

Summary. — This chapter presents several twenty-five word scales for measuring the spelling power of pupils in grades III to VIII, with complete directions for applying the scales. It provides comparative data, showing what progress in spelling power may be expected from grade to grade or from year to year. Directions given are such that many additional scales of equivalent value may be

prepared from the same sources by any one interested in measuring spelling ability.

QUESTIONS AND EXERCISES

1. What school subject has been next after spelling in its popularity as a basis for contests between schools, especially in rural districts? Give as many reasons as you can for the wide-spread popularity of this second subject in school competitions. Why have spelling contests been more popular than contests in this other subject?

2. Are all arithmetic problems equally good measures of ability to "cipher"? Why?

3. What are the different reasons you can think of that might have caused a fifth-grade class to make five times as many errors in spelling "certainly" as in spelling "weather"?

4. In what sense would it be more unfair to test spelling ability by the word "artesian" than by the word "chauffeur"? by the word "principal" than by the word "discussion"?

5. Ask your pupils to copy from your dictation the following sentences on a sheet of their usual composition paper. Do not suggest to them that they are being tested in spelling, and do not overemphasize any of the words. Each sentence may be read as often as necessary, and one phrase at a time, but no word should be read separately or by syllables.

(1) The next day he hurt his foot.
(2) My sister went to the party.
(3) At the table I shall teach you to make a chain.
(4) The second trip was very rapid.
(5) They will publish your salary in the local paper.
(6) A recent speech by a citizen started a campaign for science in the high school.
(7) This peculiar decision is quoted by those who allege that a menagerie should not be directed by a lieutenant.
(8) Without further embarrassment, the clever maneuver of the rhinoceros was easily observed from the naphtha launch.

6. Test the same pupils according to the directions given in this chapter, using Spelling List 4.

7. What percentage of your pupils spelled the words of List 4 more accurately from sentence dictation than in the spelling list? What percentage spelled exactly the same words correctly in both cases? What percentage spelled less well from dictation?

8. Which words in List 4 caused the greatest amount of difference between dictation and list spelling?

9. Why would it be unfair to compare the results of the dictation spelling test with the results in Baltimore and in the cities tested by Dr. Ayres?

REFERENCES FOR READING

B. R. Buckingham, *Spelling Ability. Its Measurement and Distribution.*

L. P. Ayres, *A Measuring Scale for Ability in Spelling.*

E. J. Ashbaugh, *Iowa Spelling Scale.*

Briggs, Kelley, and Others, "Sixteen Spelling Scales Standardized in Sentences for Secondary Schools," *Teachers College Record*, XXI, pp. 337–391.

CHAPTER VII

SPELLING SCORES OF A SCHOOL SYSTEM

COMPARISONS of the spelling scores made in different school systems, such as those presented by Figures 20 and 21 in the foregoing chapter, are useful means for discovering the relative effectiveness of the spelling instruction in the schools compared. A great mass of information must be carefully gathered and tabulated, however, before such comparisons are possible. Unless the data required for such comparative graphs are also to be used for more detailed studies of the spelling situation in each school and for possible improvements therein, it is scarcely worth while to spend the time and energy required for collecting and summarizing them. Analysis of the test results by buildings, by rooms, by recitation sections, and by other units will furnish many valuable suggestions as to methods by which the average scores may be improved. The discovery and strengthening of specific weaknesses is the most effective method of improving a general situation.

It is the purpose of this chapter to emphasize the importance of analyzing carefully the scores of a school system in spelling, or in any other test. Consideration will also be given to certain statistical processes, including the calculation of average and percentile points in distributions of spelling test scores. For these purposes

the spelling scores obtained by the white pupils of the elementary grades in the educational survey of Baltimore will be used.[1]

The continuity of the scale for spelling ability. — Any calculations with spelling test scores bring promptly to consideration the question as to the range of ability possessed by pupils having a given score in spelling. Are all pupils having a score of "10 words right" equally able spellers? Assume, for example, that in spelling the words of the first list given in the preceding chapter, a large number of fourth-grade pupils had a score of 10. Assume that all of them had spelled correctly the first ten words of the list, and that several of them had attempted to spell the eleventh word. Are the following ways of spelling this word all equally wrong: *visiter*, *wisitor*, *visetor*, *visıtor*, *vizter*, *wiztr?*

It would seem that the first and fourth misspellings were certainly less serious than the fifth and sixth, and yet, as far as their scores are concerned, all pupils who have spelled only ten words with perfect accuracy are, according to the directions, to be considered alike and given a score of 10. If the pupil who at the eleventh word was unable to do better than "vizter" seems to deserve a score of 10.4, then the one who wrote "visitor" (forgetting merely to cross the *t*) may readily be thought of as deserving a score of 10.9.

[1] The report on the spelling scores of Baltimore pupils is used here to illustrate some of the statistical methods that are commonly employed in studying and reporting educational test results. Scores in reading or in handwriting would have served equally well, perhaps. The methods are the important things to study in this chapter.

It would probably be possible to discover various degrees of error in the spelling of each word, upon the importance of which the majority of teachers could agree fairly well. For the purposes of assigning scores in this test, however, all errors are to be considered equal. Whether the eleventh word is wholly wrong or only slightly wrong, the score

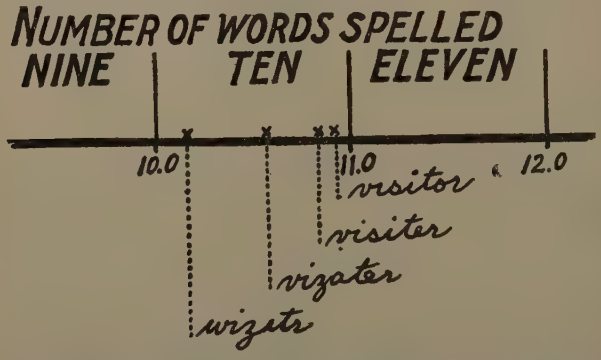


FIG. 22.— Diagram of the range of a score of ten, showing various degrees of error in spelling the eleventh word. (The scale values for the four errors are arbitrarily chosen.)

remains 10. The difference between the poorest score of 10 (no attempt to spell the eleventh word) and the best score of 10 (visitor) is much greater than the difference between the best score of 10 and the poorest score of 11 (visitor). A score of " 10 words right " applies to all who spelled only 10 words with perfect accuracy, whether they spelled an eleventh word almost right or neglected to attempt the spelling of an eleventh word.

The amount of spelling ability possessed by an individual does not increase or decrease by sudden leaps. It develops or deteriorates gradually. The scale for spelling ability is just as continuous as the scale for age or time. In the

case of the time scale, a child is not usually called ten years old until he has lived a full period of ten years, and then he is called ten for the entire time he spends living his eleventh year of life. In the case of spelling, likewise, a child has a score of 10 from the time he first succeeds in spelling the tenth word correctly up until the time he first succeeds in spelling the eleventh word accurately. The pupils who have spelling scores of 10 range in actual fact from 10.0 up to 10.999, just as the pupils 10 years old at last birthday range in actual fact from 10 years, 0 days, up to 10 years, 364.9 days.

The fact that one group of pupils is registered on 9, another group on 10, and another on 11 does not mean that all those having a score of 10 are equal, nor that all of those who have a score of 11 are a full unit better than all those having a score of 10. Spelling ability is a continuous scale, with no breaks or sudden spurts in it. The score in spelling, in age, in intelligence, or in almost any other mental test is merely a record of the last milestone passed. The average distance traveled by those between the tenth and eleventh milestones at any given moment may be assumed to be 10.5, for some of them have just barely passed the milestone at 10.0 and others are just ready to pass the marker at 11.0. In all probability there are just as many who have not yet reached 10.5 as there are who have passed beyond it.

The calculation of the median point as a summary of a distribution of spelling scores. — The spelling scores made by fourth-grade white pupils in the public schools of Baltimore were represented in Table XIV (and in Figure 20) as 9.7, a point that is a little more than one-third of the

distance from the lowest to the highest score possible in this test. This point, 9.7, represents the median, or 50 percentile point, for a distribution of more than two thousand scores, and yet only 279 pupils were actually scored as having "9 words correctly spelled." Two pupils in this grade did not spell a single one of the twenty-five words correctly, two of them spelled twenty-one words correctly, and the others ranged all the way between these two scores. It is quite evident that no single point is wholly adequate as a measure of such a large group of spelling scores, spread over such a wide range. Yet only by making summaries of many cases in a small compass can one really fix the complex situation in his mind clearly enough to compare one large group with another.

As an example of the need for brief, easily comprehended summaries of large masses of scores, the reader may try to compare with each other the five elementary grades of the Baltimore schools as represented by the distributions of spelling test scores given in Table XVI. It is fairly easy to see that there is some improvement as one passes from the fourth grade to the fifth, from the fifth to the sixth, and so on until one reaches the eighth grade. Without making summaries of the facts in each grade, however, it would be almost impossible to give any intelligible statement of just how great the improvements are.

One of the most obvious summaries to make of a group of scores is to locate some average point or central tendency for the entire group. The median or 50 percentile point is the average most commonly used in educational tests. As was seen in the case of age scores, the median point is found by going half-way into the distribution, — by

TABLE XVI

DISTRIBUTION BY GRADES OF SPELLING SCORES IN BALTIMORE, MD.

No. Words Right	Frequency of Scores in Elementary Grades				
	IV	V	VI	VII	VIII
25	—	—	—	1	3
24	—	—	1	4	21
23	—	—	2	9	27
22	—	1	3	21	42
21	2	9	14	49	100
20	1	5	33	110	148
19	4	21	55	154	153
18	8	44	120	194	144
17	16	63	159	229	149
16	17	156	240	230	145
15	56	156	225	153	87
14	87	220	194	114	53
13	120	215	191	97	29
12	161	257	141	60	17
11	227	237	89	31	3
10	230	221	70	20	3
9	279	167	31	7	6
8	263	89	23	2	—
7	240	62	5	1	—
6	126	27	5	—	—
5	86	11	—	—	—
4	66	4	1	—	—
3	26	2	—	—	—
2	18	—	—	—	—
1	9	—	—	—	—
0	2	—	—	—	—
Total	2044	1967	1602	1486	1130

going along the scale to a calculated point, on either side of which one would find an equal number of scores. In

the fourth-grade distribution of Table XVI, it would be necessary to calculate a point on either side of which 1022 scores would theoretically be found. Beginning with the low scores and counting all pupils who have scores from 0 to 8 inclusive, one would include 836 pupils, which would leave 186 (1022 − 836 = 186) to be taken from the 297 pupils with scores of 9. Now 186 out of 279 cases represents .667 of all on the step called 9, if one assumes that the 279 cases are distributed evenly over step 9. As demonstrated in the previous section, a score of 9 means anywhere from 9.0 up to 10.0. To go .67 of the distance through 9 from the lower end would bring one to 9.67, which therefore represents the calculated median point in the fourth-grade distribution of spelling scores in Table XVI.

SUMMARY OF CALCULATION

Total no. cases,	2044
50% "	1022
Total below 9,	836
To be taken in 9,	186
Total at 9,	279
Fraction of 9,	.667
Lowest point in 9,	9.0
9.0 + .667 =	9.667

The same median or 50 percentile point might be calculated from the other end of the distribution. Adding all frequencies on steps higher than step 9 gives a total of 929, which would leave 93 to be taken, from the 279 at step 9, to make up the 1022 to be passed. Assuming the even distribution of the 279 cases on step 9, taking 93 of

SUMMARY OF CALCULATION

50% of cases,	1022
Total above 9,	929
To be taken in 9,	93
Total at 9,	279
Fraction of 9,	.333
Highest point in 9,	10.0
10.0 − .333 =	9.667

them would mean going $\frac{1}{3}$ of the way down from 10.0 to 9.0. This would leave one, as in the previous calculation, at 9.667, the median or 50 percentile point for the fourth-grade spelling scores. This calculated point represents the entire group of 2044 pupils in the sense that half of them probably spelled more and half of them fewer than $9\frac{2}{3}$ words.

The arithmetic mean as a summary of a distribution. — The arithmetic mean or "average" has been used extensively in summarizing scholarship records and other facts. It is on the whole an excellent summary of the distributions obtained with carefully graduated measuring scales, such as the Standard Spelling Test Lists given in the preceding chapter. Its calculation should certainly be well understood by any one interested in the scientific determination of educational results.

The quotient obtained by adding together all of the individual measures and dividing by their number has usually been known in America as "the average," although the more exact name "arithmetic mean" has been commonly applied to it in England. The present writer, in agreement with other members of the Educational Research Association, prefers to use the word average as a general term referring to any one of several measures of the central tendency of a group of scores, among which measures the mean point and median point may be mentioned most prominently.

Of the 2044 white pupils who took the spelling test in the Baltimore fourth grades, each made a score somewhere between 0 and 21 (see Table XVI). It would have been possible to determine the arithmetic mean of these scores

without ever tabulating them, although it is easier to determine it from the frequency table than it would have been from the original 2044 individual scores. Starting with the small scores in the distribution, one sees that there were 2 pupils with a score of 0, and 9 with a score of 1. Instead of putting down two *zeros* and nine *ones* to be added to the other scores, one gets the same result by putting down $2 \times 0 = 0$ and $9 \times 1 = 9$. Table XVII gives the complete calculation of the mean point for the fourth-grade scores under discussion. The sum of all the products (in column III) of multiplying each score (in column I) by its frequency (in column II) is 19,041. As stated in previous paragraphs, however, a score of 1 means from 1.0 up to 2.0, and a score of 2 means from 2.0 up to 3.0, so that each multiplication was with the lowest possible value of each score. It is evident that each multiplication should be with the mean or central value of the score rather than with its lowest or its highest limits. Each of the 2044 cases was rated .5 of a score too low in the multiplications, and therefore 1022 ($.5 \times 2044 = 1022$) must be added to the first total of products in order to make it represent the most probably true sum of products. Dividing this sum of products, 20,063, by the total number of scores, 2044, gives 9.82 as the mean point on the scale for this fourth grade.

The chief error that has commonly been made in calculating mean points for educational measurements has been the tendency to forget that the number by which a score is called is not necessarily the central value of the score to which it applies. In the next chapter an example will be given of a scale on which each score is called by a number

TABLE XVII

CALCULATION OF THE MEAN POINT IN DISTRIBUTION OF FOURTH-GRADE SPELLING SCORES

I Score	II Frequency	III II × I	IV Central Value
0	2	0	.5
1	9	9	1.5
2	18	36	2.5
3	26	78	3.5
4	66	264	4.5
5	86	430	5.5
6	126	756	6.5
7	240	1680	7.5
8	263	2104	8.5
9	279	2511	9.5
10	230	2300	10.5
11	227	2497	11.5
12	161	1932	12.5
13	120	1560	13.5
14	87	1218	14.5
15	56	840	15.5
16	17	272	16.5
17	16	272	17.5
18	8	144	18.5
19	4	76	19.5
20	1	20	20.5
21	2	42	21.5
22	—	—	22.5
23	—	—	23.5
24	—	—	24.5
25	—	—	25.+
Total	2044	19041	

Added, to correct,
to central values 1022
Correct sum of products 20063
20063 ÷ 2044 = 9.815, the mean of all scores obtained.

which is the mid-value of the score or step, but in the case of most of the educational scales the name of the score is its lowest possible value.

Because of the danger of forgetting to correct for this error of .5 step on each individual, some readers will find it convenient in calculating mean points to multiply each frequency by the central or mean value of its score (see column IV of Table XVII) rather than multiply by the numerical name of the score and make the correction later. The added danger of errors due to multiplying by decimal fractions seems to the writer great enough in the cases of most students to warrant making the corrections later and all at once. Other readers will find it advisable to divide the first (apparent) total of products (19,041) by the total number of scores (2044), obtaining an apparent mean (9.315), to which they will then add .5 of a score to obtain the true mean (9.815). The objection to this practice is that some students fail to understand the reason for adding .5 to what appears to them to be a very accurately determined mean.

Another danger in the calculation of mean points is that in dividing one number by another, which it may not contain evenly, there is sometimes a tendency for careless thinkers to carry out the result to several decimal places and to believe in the accuracy of these decimals, regardless of the crudity of the original scales. Belief in the second and third decimal place is particularly indefensible when the first decimal, through neglect to make the correction indicated in the above paragraph, is clearly .5 too low.

The range of the middle fifty per cent as a summary. — Groups of individuals when measured on the same scale

differ from each other not only in their average scores but also in the closeness with which they are grouped about their averages. In comparing two school systems measured on the spelling lists given in the preceding chapter, for example, one might find the median scores for the one system almost exactly the same as the median scores for the corresponding grades of the other system, and yet the scores in one might be much more closely grouped about their medians than in the other system. The more closely the scores of a group cluster about their average, the greater is the evidence that pupils have been classified accurately according to their abilities. It is much easier to teach a class effectively if the pupils are approximately equal in knowledge and ability. Any report of the average scores in a test should therefore be supplemented by a statement of how closely the scores entering into each of these averages are grouped.

A brief discussion was given in Chapter V of various measures of the spread or dispersion of ages in a group of pupils. Although something can be said in favor of the total range of scores, of the middle ninety per cent, or of the middle eighty per cent, there is more to recommend the range of the middle fifty per cent as a measure of the grouping of test scores. The calculation of the limits of the middle fifty per cent involves nothing more than the determination of the 25 percentile and 75 percentile points in the distribution. These points, whether in a distribution of age scores or spelling test scores, are determined in a manner exactly comparable to the determination of the median or any other percentile point.

The 25 percentile (or first quartile) point in the fourth-

grade distribution of spelling scores given in Table XVI would be found by locating the point on the scale below which 511 of the 2044 pupils were rated. By the time one has counted all the cases with scores less than 7, he will have included 333 of the desired 511 cases, and 178 others will have to be taken from the 240 at 7. Assuming that these 240 are distributed evenly along the scale from 7.0 up to 8.0, to take $\frac{178}{240}$ of them would carry one .74 of the distance. The desired 25 percentile point is therefore at 7.74 words.

SUMMARY OF CALCULATION

25% of cases,	511
Total below 7,	333
To be taken in 7,	178
Total at 7,	240
Fraction of 7,	.742

7.0 + .742 = 7.74

The 75 percentile (or third quartile) point in this fourth-grade distribution would be found by locating the point below which 1533 of the 2044 pupils were rated. There are 1345 cases below 11, which leaves 188 to be taken from the 227 at 11 in order to make up the desired 1533. These 188 cases probably cover the lowest .828 (188 ÷ 227 = .828) of the part of the scale called 11, which would place the 75 percentile at 11.83 words.

SUMMARY OF CALCULATION

75% of cases,	1533
Total below 11,	1345
To be taken at 11,	188
Total at 11,	227
Fraction of 11,	.828

11.0 + .828 = 11.828

The limits of the middle fifty per cent of fourth-grade spelling scores in Baltimore were therefore 7.74 and 11.83. The range of the middle fifty per cent (the difference be-

tween these two limits) was 4.09 words. The report that the median pupil in this grade spelled correctly 9.7 words and that the range of the middle half of the scores was from 7.7 to 11.8 words is a much more complete statement of the total situation than the mean or the median point alone would be.

The calculation of the limits of the middle fifty per cent, or of some other measure of the closeness of the grouping in a distribution of scores, assists in answering such a question as that of whether the pupils in the upper grades are more (or less) accurately classified according to spelling ability than are the pupils in the earlier grades. Are fourth-grade pupils more alike in spelling ability than sixth- or eighth-grade pupils? Table XVIII gives the median and mean points and the limits of the middle fifty per cents of the distributions (given in Table XVI) for the five upper elementary grades in Baltimore.

TABLE XVIII

AVERAGES AND DISPERSION OF SPELLING SCORES IN ELEMENTARY GRADES

SCHOOL GRADE	IV	V	VI	VII	VIII
CENTRAL TENDENCY OR AVERAGE					
Mean point	9.8	12.7	15.0	17.0	18.4
Median point	9.7	12.6	15.2	17.1	18.5
DISPERSION					
25 percentile	7.74	10.59	13.19	15.26	16.58
75 percentile	11.83	14.83	16.94	18.88	20.40
Middle 50% range	4.09	4.24	3.75	3.62	3.82

The last line in Table XVIII shows that the Baltimore eighth grade was slightly more homogeneous, more uniform in terms of units on this spelling scale at least, than the fourth grade, although the seventh grade seems to be a bit more compact than either, and the fifth grade more widely dispersed than any other grade. Figure 23 shows that the

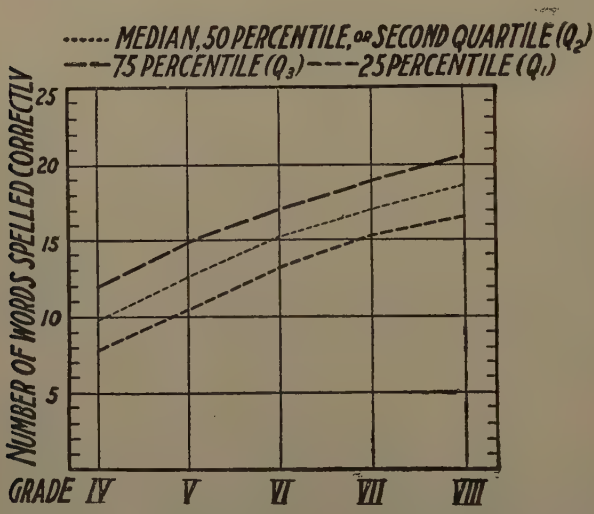


FIG. 23. — Range of the middle fifty per cent of spelling scores in each grade of the Baltimore public schools.

amount of difference between grades in the range of the middle fifty per cent is very small. The elimination from school of many poor spellers at the end of the fifth grade, and more careful selection of pupils for promotion into the upper grades, may have contributed to the slightly more satisfactory measures of dispersion in the higher grades.

The same questions might be asked with regard to the age groups. Are ten-year-old pupils more nearly alike

TABLE XIX

DISTRIBUTION BY AGES OF SPELLING SCORES IN GRADES IV TO VIII

No. Words Right	Frequency of Scores at Age									
	8	9	10	11	12	13	14	15	16	17
25	—	—	—	—	1	2	1	—	—	—
24	—	—	—	—	3	13	6	3	1	—
23	—	—	—	1	9	17	6	3	1	1
22	—	—	1	5	9	21	22	8	1	—
21	—	3	4	18	25	57	44	16	5	—
20	—	1	5	30	59	87	74	30	9	1
19	—	2	14	40	78	137	81	30	3	1
18	1	5	28	69	134	125	100	27	14	1
17	1	8	40	81	147	184	88	47	11	—
16	—	14	76	131	194	189	108	53	12	2
15	5	25	85	125	175	141	64	40	9	2
14	10	38	102	144	145	133	61	21	5	1
13	6	64	88	135	157	121	40	22	3	—
12	9	53	122	133	145	99	45	19	2	1
11	18	76	113	133	101	81	36	13	2	—
10	11	77	131	87	108	81	28	9	2	—
9	5	92	120	88	88	55	26	11	2	—
8	12	63	96	73	61	39	18	6	1	—
7	10	51	83	58	38	36	15	6	—	—
6	3	35	41	35	21	15	4	3	—	—
5	1	10	23	22	27	6	1	2	—	—
4	2	14	9	23	9	8	1	—	—	—
3	—	8	6	3	8	—	1	—	—	—
2	1	2	9	2	—	2	1	—	—	—
1	—	—	4	1	1	—	1	—	—	—
0	—	—	1	—	1	—	—	—	—	—
Total	95	641	1201	1437	1744	1649	872	369	83	10
Median	11.1	10.6	11.7	13.4	14.7	16.0	16.9	16.6	17.3	16.5

in spelling ability than fourteen-year-old pupils? Do pupils become more or less alike in ability as one considers older groups? Table XIX gives the distributions of spelling scores by ages for Baltimore pupils in Grades IV to VIII. From this table it is possible to calculate the range of the middle fifty per cent for each age group, in the same manner as this measure of dispersion was calculated for each grade from the facts of Table XVI. The limits and range of the middle fifty per cent for each age are given in Table XX.

TABLE XX

DISPERSION OF SPELLING SCORES IN BALTIMORE BY AGES (GRADES IV–VIII)

AGE	8	9	10	11	12	13	14	15	16	17
Range of middle 50 per cent	4.36	4.36	5.30	5.50	5.46	5.47	5.19	4.86	3.53	4.25
25 percentile	8.56	8.64	9.24	10.62	11.74	12.90	14.01	14.06	15.42	15.25
75 "	12.92	13.00	14.54	16.12	17.20	18.37	19.20	18.92	18.95	19.50

Eleven-, twelve-, and thirteen-year-old groups of pupils at Baltimore were very much alike in the variabilities shown by the distributions of their scores, each group requiring a range of about 5.5 words to include the middle half of all scores. The ten-year-old group required a range of 5.3 words and the fourteen-year-old group took 5.2 words as a middle fifty per cent range. The other age groups are incomplete, as can be seen by observing (Table XIX) the total number of pupils at each age; this incompleteness should be expected from the fact that only

pupils in grades IV to VIII were included. It is doubtful whether the amount of dispersion found for the ten-year-olds is even approximately correct for the general ten-year-old situation at Baltimore, for there must have been in the second and third grades, many pupils of this age who were not tested. At the other end of the line there were many fourteen-year-old pupils in the high schools, where no tests were held. The range of the middle fifty per cent of scores for pupils fourteen years old or older and for pupils ten years old and younger must, therefore, be considered as inadequately determined.

Amount of overlapping as means of comparing groups. — When one wishes to make comparisons between two or more groups of pupils measured on the same scale, the most common practice is to state the median or mean score of each group and then to give some measure of the variability or dispersion of the scores in each group. It is almost never satisfactory to give only an average score, except where the persons to whom the report is being made are thoroughly familiar with the scale and know the variability without being told. If possible, the two distributions under consideration should be compared in detail. Frequently it will be found helpful to plot the surfaces of frequency upon either the same base, or upon the same page with equivalent base lines, as in Figure 24.

The facts presented by Figure 24 are exactly the same as those previously furnished by Table XVI. Under some conditions it would have been better to have converted each *number* of pupils having a given score into a *percentage* of the total number of pupils in the group, in order to make the surface for the eighth grade exactly the same in area

as the surface for the sixth grade or for the fourth grade. The writer has found it valuable for purposes of clear

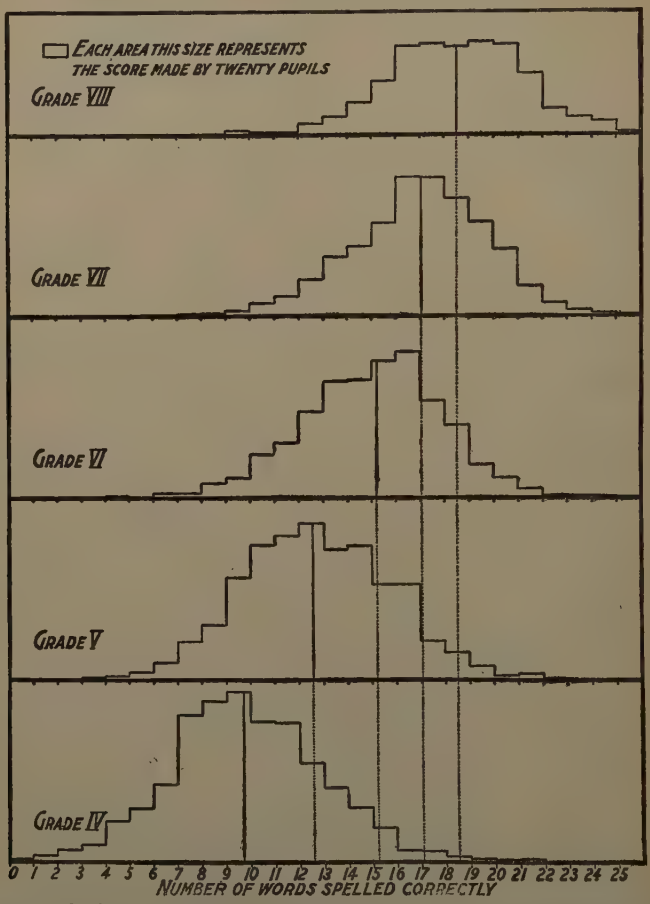

EACH AREA THIS SIZE REPRESENTS
THE SCORE MADE BY TWENTY PUPILS
GRADE VIII
GRADE VII
GRADE VI
GRADE V
GRADE IV
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
NUMBER OF WORDS SPELLED CORRECTLY

FIG. 24. — Surfaces of frequency representing the distributions of spelling scores in each grade from IV to VIII at Baltimore, Md. Perpendicular lines, erected at the median points, divide each area into two exactly equal parts. These lines extended downward show just what part of the area for Grade IV is above the median of each other grade.

interpretation, however, in those instances where elimination or other large external factors have been at work on one or more of the groups, to represent actual numbers of cases rather than percentages. In Figure 24, therefore, the altitude of a surface at any point represents the actual frequencies of scores, while the distance to the right represents the size or excellence of the scores. The reader should note especially the general drift of the peak toward the right as his attention passes from lower to higher grades.

Although it is possible to see progress in spelling from grade to grade in such a chart as Figure 24, it would be difficult to describe the situation adequately without the presence of the actual drawing. The usual statements of median or mean scores with amounts of dispersion are not always adequate. Another way of comparing groups has, therefore, been developed, that of telling how much of one group is as good or better than the median of the other group. What fraction of the fourth-grade distribution in Figure 24 (or in Table XVI), for example, is higher than the median point in the fifth grade, and what fraction of the fourth-grade distribution surpasses the median of the eighth grade? Definite answers to these questions should be helpful to any one interested in the amount of difference between these two grades.

The calculation of the amounts by which one distribution overlaps another can be made by any one who has learned the calculation of a percentile value. The fifth-grade median, for example, was found to be 12.64 words. There were 161 pupils (see Table XVI) in the fourth grade with a score of 12 words, which means of course from 12.0 up to 13.0 words. Sixty-four hundredths of these 161 were

probably below the median of fifth-grade pupils, but the other thirty-six hundredths were above. Now .36 of 161 pupils (or scores) would be 57.96. There were then in the fourth grade 57.96 pupils with a score of 12 who may nevertheless be assumed to have a better score than the median fifth-grade pupil. In addition, there were 120 fourth graders with a score of 13, 87 with a score of 14, and so on. A total of 311 pupils in the fourth grade had 13 or more words, and with the 57.96 pupils who may be assumed to have better than 12.64 words, there were in all 368.96 pupils in the fourth grade doing better than the median fifth-grade pupil's score. These 368.96 make up 18.05% of the total fourth grade. We may therefore say that 18% of the fourth grade did as well or better than the typical fifth-grade pupil.

As to the overlapping of the fourth grade on the eighth, one finds the eighth-grade median point at 18.51 and that 8 pupils had scores of 18 in the fourth grade. One may assume that .49 of the 8, that is 3.92 pupils, made a score higher than 18.51. There were also 7 fourth-grade pupils with 19 or more words, so that altogether 10.92 of the 2044 fourth-grade pupils, which is approximately .5 of 1% of them, did as well as the typical eighth-grade pupil in spelling. This method of describing the differences between groups in terms of the percentage of one group that equals or exceeds the median of the other group is thoroughly scientific, is often very practical, and is easily understood.

Percentile tables and curves.—Where it is not desirable to reproduce tables giving the entire distributions of scores, and where reports of the central tendency and variability seem inadequate representations of measured situations,

percentile curves are frequently very useful. Table XXI and Figure 25 represent the same spelling scores of

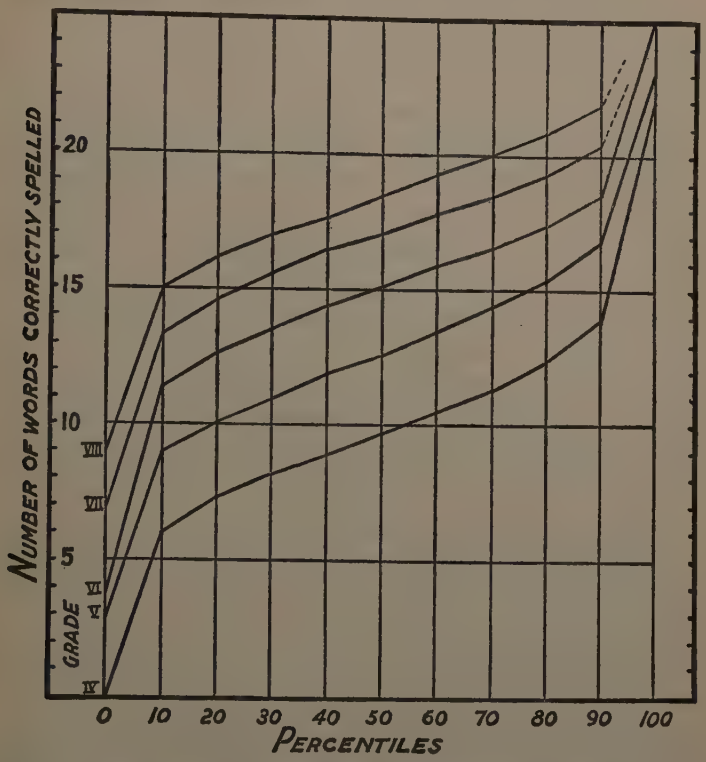


FIG. 25. — Decile curves for distributions of spelling scores in Grades IV to VIII of the Baltimore public schools. The data for these curves are given in Table XXI and were calculated from the distributions given in Table XVI.

the white pupils of Baltimore as were represented in Table XVI and in Figure 24.

One advantage of percentile curves is found in the effectiveness with which they break up the easily obtained

idea that a mean or median score in a group is really typical of all pupils in the group. If a Baltimore child were to be found with a spelling score of 15, one could not safely say that this pupil belonged in the sixth grade, although 15 is a typical score for sixth-grade pupils. He might be a member of any grade from the fifth to the eighth. If he were placed in the fifth grade the chances are, according to the facts of Table XXI, that he would be among the best 20 per cent of spellers in his class; but if placed in the eighth grade he would probably be in the lowest 20 per cent of spellers in his class. He might, however, feel at home in either class. The calculation and plotting of percentile values help the young investigator to avoid the easily developed habit of thinking that a pupil, who makes a score which is the average for a given grade, must therefore be transferred at once to that grade. A certain

TABLE XXI

DECILE POINTS IN SPELLING SCORES FOR ELEMENTARY GRADES

BALTIMORE SPELLING IN GRADES	IV	V	VI	VII	VIII
100 percentile	22.0	23.0	25.0	25.0	25.0
90 "	13.9	16.7	18.6	20.4	21.8
80 "	12.4	15.4	17.4	19.3	20.8
70 "	11.3	14.4	16.6	18.5	20.0
60 "	10.5	13.5	15.9	17.8	19.3
50 "	9.7	12.6	15.2	17.1	18.5
40 "	8.9	11.9	14.4	16.5	17.7
30 "	8.2	11.0	13.6	15.7	17.0
20 "	7.3	10.1	12.7	14.7	16.2
10 "	6.0	9.0	11.4	13.3	15.0
0 "	0	3.0	4.0	7.0	9.0

amount of overlapping of one grade on the next is unavoidable and not wholly undesirable.

The effect of age on the scores of pupils in the same grade. — Not all pupils take the same number of years to reach a given school grade. It is presumably the older pupils in a grade who have taken more time to reach the grade and who have less ability to do the work. The extent to which this was true of spelling ability in Baltimore is shown by Table XXII, which gives the median point in

TABLE XXII

MEDIANS OF SPELLING SCORES BY AGES WITHIN EACH GRADE

AT AGE	8	9	10	11	12	13	14	15	16	17
Grade IV	11.1	10.2	9.6	9.3	9.2	8.9	9.5	7.8	—	—
" V	—	13.9	13.5	13.0	12.2	11.2	11.2	10.6	—	—
" VI	—	—	16.1	15.8	15.2	14.5	14.3	13.9	11.5	—
" VII	—	—	19.0	18.3	17.3	17.1	16.9	16.4	15.7	—
" VIII	—	—	—	19.5	18.7	18.9	18.6	17.8	17.8	17.5

the distribution of spelling scores for pupils of each age in each grade. This table and Figure 26 give definite support to the statement that it is the younger pupils in each grade who are being retarded and held back in their progress through the school grades. Older pupils have been promoted beyond their ability to do the work, while younger pupils have been held in grades lower than their abilities would suggest.

Such analysis of scores as is presented in Table XXII and in Figure 26 is made possible by the use of exactly the

same test list or scale in all grades and schools. If lists of varying difficulty had been used in different grades, it would not have been so easy to show the effect of the ages of pupils on their scores. It is worth while, of course, to measure achievement with any objective scale, but it is most worth while to measure it with such a scale as will enable one to make as complete an analysis as possible of the situation discovered.

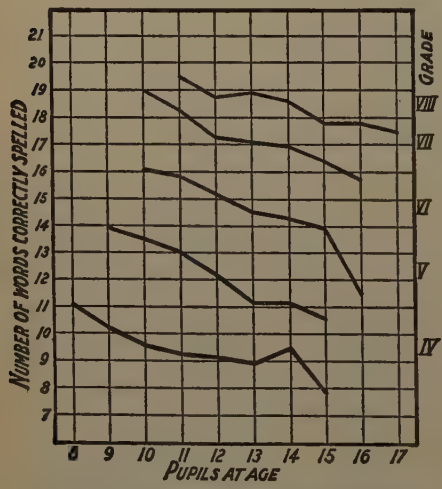


FIG. 26. — Showing that older pupils in the same grade have lower abilities. Median scores in spelling by age within each grade.

It has been customary to compare one school with another on the basis of the median scores obtained in their corresponding grades. This is not entirely fair to pupils and teachers in all instances. A study of Figure 26 suggests the possibility that a school might get a very high median score in its eighth grade, for example, by holding its slower pupils in the lower grades or by allowing them to drop out of school altogether. Another school, even though its pupils and teachers had done better work, might have a lower median score in the eighth grade, just because pupils had been kept in school and promoted regularly. Public schools should not be commended for their efficiency in getting slow pupils to drop out of school work,

and yet that is exactly the way in which some schools are "maintaining high standards."

The comparison of pupils of the same ages in different schools is better than the comparison of pupils in the same grades, inasmuch as the age-for-age comparison does not give an unfair advantage to the school that holds its slow pupils in the lower grades. Within the compulsory age limits, the relative efficiency of a school can be measured in part by what it has done for pupils of a given age. Even the age-for-age comparison is not entirely just, for pupils in one community may enter school somewhat earlier in life than the pupils in another community. Beyond the age when pupils are required by law to attend school, the age-for-age comparison is influenced by relative amounts of elimination as well as by relative efficiency of the teachers. Furthermore, the levels of intellectual ability in two communities may differ considerably, in which case the school having the most efficient teachers of spelling may actually have the lowest spelling ability. Assuming that two schools contain pupils of equal intellectual powers, however, comparisons should be made between pupils of the same age as well as between pupils of the same grade. The tabulation of scores by age within each grade supplies the data for both comparisons and shows at the same time some significant things regarding the effects of the promotion scheme used in the school system.

Summary. — This chapter shows some of the ways in which the test scores of a school system may be treated statistically in order to reveal the important general facts. It is assumed that in any school system comparisons will be made among the results of different rooms of the same

school, among the achievements of different schools, and among the records of different types of schools in the same city. Emphasis is placed in this chapter upon the value of analyzing the scores of a school system by age within each school grade as a basis for age and grade comparisons. Attention is directed again to a common error in the calculation of an arithmetic mean for educational test scores, and instructions are given for the calculation of percentile points, including the median or 50 percentile point, from tables of frequency. Comparisons of groups by means of surfaces of frequency, percentile curves, and amounts of overlapping are illustrated from the spelling scores of the white pupils tested in 1920 for the educational survey of Baltimore, Maryland. The range of the middle fifty per cent of scores is used in this chapter as a measure of the dispersion or homogeneity of the groups measured.

QUESTIONS AND EXERCISES

1. If you were required to assign fractional spelling scores, giving a full credit of 1.00 for a perfectly spelled word and credits of from .99 down to 0 for each incorrectly spelled word (the amount of credit depending in each case upon how badly misspelled the word might be), how much fractional credit would you assign each of the following attempts: lessen? lesin? lesn? secretery? secatary? sec.? athel-etic? athletick?

2. Calculate the median or 50 percentile point for each of the grade groups tabulated in Table XVI, carrying out each one to the third decimal place. How do your calculations compare with the median points given in Table XVIII? What rules seem to have been used in dropping the second and third decimal places?

3. Calculate the mean point for each age group tabulated in Table XIX. How closely do these correspond with the median points given at the bottom of the table?

4. Which grouping is spread over the larger number of scale units, that by grades or that by ages? What was the approximate average range in Baltimore for the middle fifty per cent of a grade? of an age group?

5. Make a statistical comparison of the distribution of spelling scores for sixth-grade pupils in Baltimore with the distribution for twelve-year-old pupils. Represent the facts graphically as well as numerically.

6. Prepare percentile curves for the distributions by ages given in Table XIX. Why do the eight-, nine-, fifteen-, sixteen-, and seventeen-year-old curves differ so much from the others?

7. Why do older pupils tend to spell more poorly than younger pupils in the same grade? Will the same amount of drill in spelling help the younger pupils more or less than it helps the older? Why?

8. How do the scores in spelling of your own school system compare with those found in Baltimore?

REFERENCES FOR READING

Report of the Survey of the Public School System of Baltimore, Maryland, Volume 2, pp. 218–244, and 322–346.

Report of the Survey of the Public School System of Hackensack, New Jersey, pp. 97–101.

Report of a Survey of the School System of St. Paul, Minnesota, pp. 427–488.

H. O. Rugg, *Statistical Methods Applied to Education*, pp. 110–120.

CHAPTER VIII

A TEST FOR HANDWRITING QUALITY

HANDWRITING has long been recognized as one of the important arts to be learned at school. Two objectives have stood out prominently in the instruction offered. One has concerned the artistic and the other the practical value of handwriting. During certain periods in the history of education the emphasis of writing instructors was on the beauty and perfection of their pupils' written products. At the present time handwriting is considered chiefly as a tool for the expression of ideas.

The invention and wide-spread use of the printing press has undoubtedly contributed to the dwindling of interest in the artistic side of handwriting. As long as books were rare and were the prized possessions of a limited few, it was natural that efforts should be made to give each book an individuality and some artistic value. As soon as printed books became plentiful, the demand for them almost completely superseded that for the more highly illuminated hand-written books which had previously been the only ones obtainable.

There is now a general tendency to consider legibility as the first objective to be sought for and maintained in the teaching of handwriting. After a pupil has learned to write legibly, a second objective to be sought is that the pupil shall write rapidly enough to keep the mechan-

ical process from retarding or interfering seriously with the development of his thoughts. Modern instruction in handwriting therefore recognizes two elements in writing efficiency: speed and quality. Some pupils write slowly but well, and others write rapidly but poorly, while the ideal is that all should write rapidly and well.

This chapter will concern itself more directly with measurements of handwriting quality than with tests of handwriting speed. It should be noted, however, that neither of these two elements can be used as an index of writing efficiency when the other is ignored. Speed has a direct influence on quality, and quality on speed. Nevertheless, speed of writing will be considered less fully in the following pages than the quality of the handwriting produced.

Directions for a standard test of handwriting quality.[1] — See that each pupil is provided with a sheet of the usual composition paper, pen, ink, and (if convenient) a blotter.

While the papers are being distributed, write one of the following topics on the blackboard where all pupils can see it.

The Saddest Event of My Life.
How I Learned a Lesson.
The Funniest Thing I Ever Saw.
My First Lie.[2]

Use care to copy the topic so legibly that all in the room can easily read it. As soon as the papers have been distributed and all the pupils are ready to write, call attention to the topic on the blackboard as follows:

"I have written on the blackboard the topic for a story. I want

[1] A discussion of the reasons for the specific directions given will follow rather than accompany these instructions.

[2] As far as the handwriting is concerned, it makes little difference which of these four topics is chosen. There is a difference, however, in the quality of the English compositions that will be written on the different topics.

you to write as interesting a story as you can on this topic. You may make up a story, if you need to, but I think it would be better to write about something that really happened.

"You will have twenty minutes in which to write. Do not ask any questions or get any help from any one else. Write carefully. If you make a mistake, correct it at once. Keep your paper as neat as possible. Use your best handwriting. You may write on both sides of the sheet if you need to, but I would rather have the whole story on the first page, if possible.

"Do you wish to ask any questions before you begin?" Pause for a few seconds and answer any questions about the assignment. If a question is asked about the quality of handwriting to be used, remind the pupils that you have told them to write as well as possible, but do not emphasize this point. When it is clear that the assignment is fully understood, tell the pupils to begin their stories.

Allow exactly twenty minutes after telling the pupils to go to work, and then say, "The time is up. Stop writing. Lay aside your pens." See that all pupils stop writing at once. Give the next instructions, carefully pausing after each item and if necessary demonstrating where the pupils are to write.

"Now I want a record of who you are, what grade you are in, and how old you are. Write your name in the upper right-hand corner of the first page of your story. Just under your name, tell the grade to which you belong. Write your age at your last birthday just under the grade. First, your name; then your grade; and then how old you were on your last birthday." If other information is desired, such as the name of the country in which the child's parents were born, the instructions should be given in simple but explicit terms.

Obtaining a sample of a child's handwriting. — The directions given above for a "standard test of handwriting quality" are also the directions for an English composition test. These directions were used in order that the handwriting should be as nearly as possible the quality commonly used by the pupil in his school work outside the writing recitation. Many pupils have a different

script for their writing teachers from that which they use for their teachers of English, of history, or of mathematics. Pupils usually write one or two degrees higher, as measured by the Thorndike Scale, if they realize that handwriting is the element to be observed, although there are a few who do not write as well when making special efforts as they do when thinking of something else. The present writer therefore prefers to measure the quality of handwriting used by pupils in writing English compositions or friendly social letters rather than the quality of specially prepared handwriting exhibits. The practical value of instruction in writing should be indicated by the quality of the writing pupils do in their other classes.

When one has obtained a written composition from each pupil, the problem at once arises as to which lines of the composition are to be taken as typical of the child's usual handwriting. The first few lines are frequently much more carefully written than the remainder of the composition and may not therefore serve in all cases as typical performances. The last few lines, on the other hand, are often written in such a hurry as to be below the general average in quality. The most adequate sample may usually be taken from the middle of the composition. To make certain of having a typical sample of each pupil's writing, the examiner may before rating any papers run through the entire pile, checking with a red pencil two or three adjoining lines that seem typical in each pupil's paper. The writer suggests that lines 10, 11, and 12 be used as the basis for rating each pupil's handwriting, unless these lines happen in any particular composition to be poor representatives of the quality of writing used in all other parts of the composition.

Having marked the lines by which each pupil's work is to be judged, the examiner may begin rating the quality of handwriting in these lines. The rating process should be carried on with all of the systematic care suggested in the next section of this chapter. As soon as a child's writing has been compared with the samples of the scale, and has been finally rated as nearer in quality to some one than to any other scale value, a record of the rating should be made on the child's paper. If other judges are to rate the same paper, the record of the judgments should be written on the back of the composition, in order that ratings previously assigned may not influence later judges. If only one rating is to be given each sample, the record may as well be made at the top of the page, near the pupil's name.

Teachers who have not been trained in the use of the scale should not be satisfied with only one rating of the handwritings of their pupils. It is desirable to have independent ratings by at least three different persons on each paper, although one good rating is often superior to two or three carelessly assigned ratings. Where three independent ratings of the same paper do not agree, the middle rating of the three may be considered the most nearly correct. If two of the three independently assigned ratings are identical, the value on which they agree may be considered as the final score for the paper. For convenience and speed in tabulation, it is desirable that the final score of a paper be clearly indicated on the face of the composition somewhere near the name and age of the writer.

Directions for rating the quality of each pupil's handwriting. — The first step in learning to rate the quality of handwriting is to become thoroughly familiar with the

scale to be used. The Thorndike Scale[1] for measuring the quality of handwriting consists of a large sheet of paper, about two feet square, on which printed samples of handwriting are arranged in four horizontal rows, the poorest sample being at the left end of the top row, and the best sample being at the right end of the bottom row. The first or top row represents five different degrees of merit, which are numbered 4, 5, 6, 7, and 8. The second row contains three somewhat more legible series of samples, numbered 9, 10, and 11. The third row contains samples of qualities 12, 13, and 14; and the fourth row contains qualities 15, 16, 17, and 18.

The amount of difference in quality between any two consecutively numbered samples on the Thorndike Scale is approximately one-tenth of the difference between the best and the poorest writing by pupils in the upper elementary school grades. In so far as a considerable number of competent judges were able to estimate, the difference between quality 8 and quality 9 in the scale is equal to the difference between quality 12 and quality 13, or to the difference between any other two consecutively numbered qualities. In the same sense, the difference between qualities 10 and 14 is twice as great as the difference between qualities 8 and 10, and four times as large as the difference between qualities 14 and 15. It should be noted that the differences indicated on this scale are differences in "quality" or "merit."

[1] The Thorndike Handwriting Scale may be purchased from the Bureau of Publications, Teachers College, Columbia University. The derivation of this scale is described in Chapter XX of this volume and also in the *Teachers College Record* for March, 1910.

Figure 27 gives a general idea of the nature of the Thorndike Scale. Alternate steps of the scale are represented in Figure 27, and only one sample is given to represent each step. Each quality on the most frequently used parts of the scale itself is represented by two or more samples, unlike in slant and general character. One who desires to use this scale should secure a copy of the original full-sized scale issued by the publishers.[1] Figure 27 is intended merely as an introduction to the nature of the scale for those who are not familiar with it.

To measure the quality of any pupil's handwriting, one should place it on the sheet containing the samples of the Thorndike Scale and compare the quality of the writing with the quality of each printed sample that seems to have approximately the same merit. The quality of the written sample may be assigned the numerical value of the printed sample whose merit most nearly matches the merit of the written lines. Care should be taken to avoid confusing a mere likeness in the slant or quality of the lines with equality of general merit. One who is expert in the use of the scale may wish to assign values intermediate between those on the scale, giving such ratings as $11\frac{1}{4}$, 10.5, or 12.8. For all practical purposes, however, it is better to assign unit values only, giving to each writing the score of the printed sample which most nearly equals it in merit.

[1] It would also be advisable for one who plans to make extensive use of this scale to secure from the Bureau of Publications at Teachers College a copy of the booklet entitled *Teachers Estimates of the Quality of Children's Handwriting*, by E. L. Thorndike. This booklet gives samples of handwriting and directions for the practice and improvement of one's ability as a judge of handwriting.

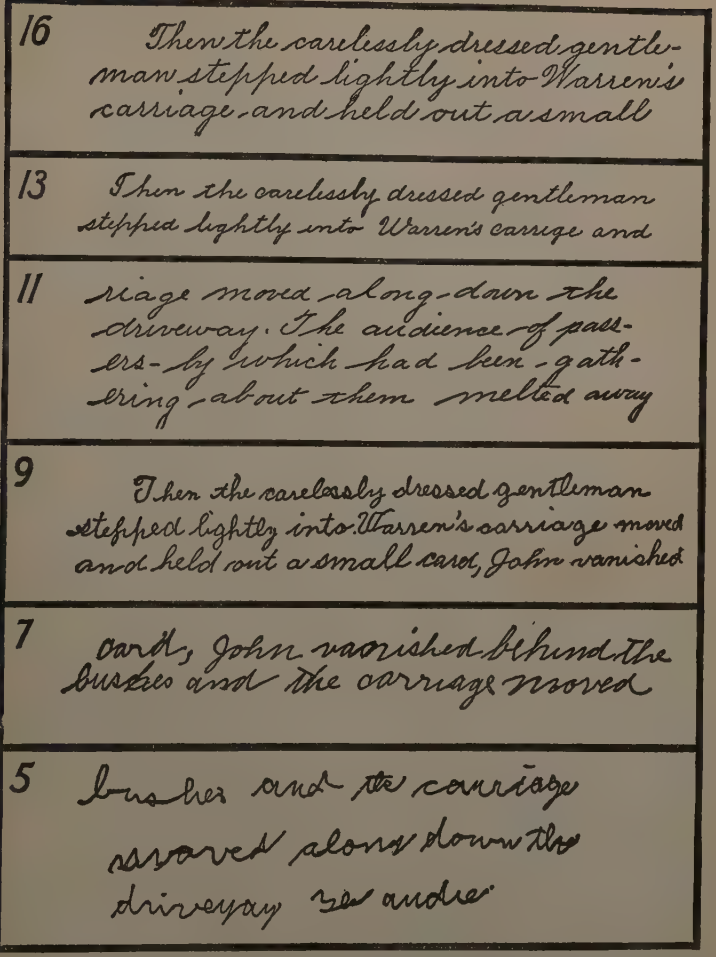

16
Then the carelessly dressed gentleman stepped lightly into Warren's carriage and held out a small
13
Then the carelessly dressed gentleman stepped lightly into Warren's carrige and
11
riage moved along down the driveway. The audience of passers-by which had been gathering about them melted away
9
Then the carelessly dressed gentleman stepped lightly into Warren's carriage moved and held out a small card, John vanished
7
card, John vanished behind the bushes and the carriage moved
5
bushes and the carriage
moved along down the
driveway

FIG. 27. — Selected samples from the Thorndike Handwriting Scale. In the scale itself, which ranges from quality 4 to quality 18, each step is represented by several samples. The average elementary school graduate uses a quality little better than 11 in his usual writing.

The grouping of values which results from this plan of assigning scores is represented in Figure 28. If a child's writing is nearer to quality 9 than it is to quality 10 or to quality 8, — that is, if it is between 8.5 and 9.5 — it is to be called 9, whether its true value is 8.6, 9.0, or 9.49. In the few cases where it seems impossible to judge whether a child's writing should be rated 9 or 8, the matter may as well be settled by tossing a coin or by some other chance method, for the value 8.5 should be grouped half of the time with 8 and the other half with 9.

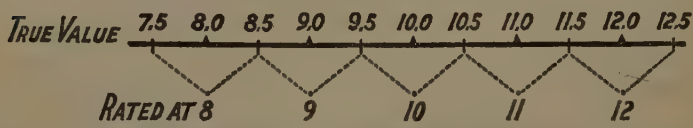


FIG. 28. — Quality of handwriting rated to nearest sample on the scale. Note that "quality 10" would include everything from 9.5 to 10.5.

In assigning scores to samples of writing, one should adopt a systematic plan for making the comparisons. It has been found that most people, when they begin with the low end of the scale in making comparisons, tend to assign values that are too low. On the other hand, when they begin their comparisons at the high end of the scale, they tend to assign high values. One tends apparently to be too easily satisfied with the identity of amounts in such qualities as handwriting merit. The best practice is to pass each written sample up the scale until a step is found that is decidedly more meritorious than the sample being rated, then to reverse the direction and to come down the scale from the high end until a step is found which is less satisfactory than the one being measured. Somewhere between these two steps on the scale a compromise value

will be found which can be identified as more nearly equal in merit than any other to the sample being rated. It would be wise to come down the scale before going up about as often as one goes up before coming down, in order to avoid any constant error that might result from always following the same order.

Standard scores on the test of handwriting quality. — The scores that have been obtained in four city school systems, where comparable measures of the quality of handwriting have been obtained, are given in Table XXIII.[1]

TABLE XXIII

QUALITY OF HANDWRITING IN FOUR CITIES BY GRADES (THORNDIKE SCALE)

City	Median Quality in Grade				
	IV	V	VI	VII	VIII
Baltimore, Md.	8.5	9.3	9.7	10.6	10.6
Des Moines, Iowa	8.1	8.4	8.9	9.5	10.0
Paterson, N. J.	8.2	8.7	9.4	9.8	12.7
Hackensack, N. J.	7.7	7.9	9.0	11.8	12.1

Only between Hackensack and Baltimore is it possible as yet to compare the results of this handwriting test by ages. It must be recalled, in comparing these two cities, that the third grade was included at Hackensack, which procedure would supply lower eight- and nine-year-old scores than were obtained in Baltimore, where all pupils

[1] The quality of handwriting has been measured in many other city school systems, but the samples measured were not in the "usual writing" found in English composition or history papers.

of these ages came from the fourth grade or above. It should also be noted that the high school scores were included at Hackensack, which would supply a higher selection of fourteen-, fifteen-, and sixteen-year-old pupils than

TABLE XXIV

QUALITY OF HANDWRITING IN TWO CITIES BY AGES
(THORNDIKE SCALE. USUAL WRITING)

CITY	MEDIAN QUALITY AT AGE								
	8	9	10	11	12	13	14	15	16
Baltimore, Md.	8.4	8.6	8.8	9.3	9.8	9.8	9.8	9.8	10.4
Hackensack, N. J.	8.0	8.1	8.2	9.2	10.9	11.7	12.7	12.3	13.2

was used at Baltimore, where only elementary school scores were considered.

Influence of test directions on the quality of handwriting. — The instructions given in an earlier section of this chapter seek to secure the normal, every-day quality of writing used by the pupils. It has been much more customary, when measurements of handwriting were to be made, to try to secure a sample of the pupil's best writing. Somewhat less frequently the effort has been to have the pupil write as rapidly as possible and as well as possible, both at the same time. The handwriting scores that have been published in school surveys and superintendents' reports have frequently overlooked entirely these differences in the methods of obtaining samples to be rated. It is obviously unfair to compare the best handwriting in one

city with the most rapid writing in another and with the usual writing in still another, and yet this kind of comparison has often been made.

Some years ago the writer undertook to discover what the effect of these different types of test instructions would be on the same classes.[1] In several schools of suburban New York the pupils were first asked to write a composition, the directions being approximately those given earlier in this chapter. They were then given another sheet of paper on which the following letter had been mimeographed:

"Dear friend:

This is a line of my best handwriting. I shall write another at the close of the term. Please tell me then how much I have improved.

Your friend,"

The pupils were told to copy this letter as well as they possibly could.

For a third test, the pupils were asked to repeat the sentence, "Mary had a little lamb." They were then informed that the object of the test was to see how rapidly they could write. The following instructions were given:

"Hold up your pen. When I say 'Begin,' write 'Mary had a little lamb' just as many times as you possibly can. As soon as I say 'Stop,' you must put your hand up again, even if you are in the middle of a word. Write as rapidly as you can and yet do it as well as you can."

As soon as the test was thoroughly understood, the word "Begin" was spoken and the pupils were allowed to

[1] See pp. 201–209 of the writer's *Report of a Survey of Public Education in Nassau County, New York.*

write for exactly two minutes, at the end of which they were stopped and the papers collected.

Each sample of handwriting was rated independently by three different judges using the Thorndike Scale according to the instructions given earlier in this chapter. The median quality of each type of writing is given in Table XXV for the seventh and eighth grades in three school systems. The reasons for the differences in the effects

TABLE XXV

MEDIAN QUALITY OF HANDWRITING UNDER VARYING CONDITIONS

School System	Seventh Grade			Eighth Grade		
	Ordinary Writing	Best Writing	Rapid Writing	Ordinary Writing	Best Writing	Rapid Writing
X	9.5	12.1	9.5	11.1	11.6	11.0
Y	8.4	9.9	9.8	10.3	9.8	10.2
Z	7.0	9.3	10.4	7.8	9.7	10.9

of the instructions on the pupils of the three systems were easily found by observing the writing instruction offered in each system. The writing supervisor in System X almost never called on the pupils to write rapidly. Her whole emphasis was on the quality of handwriting done. The pupils in this system were so painfully deliberate in their writing as to give the impression that they were drawing the letters rather than writing. The fact that writing was a special " stunt " for them probably accounts for their poor showing in ordinary writing and in rapid writing. In System Z, on the other hand, the writing

supervisor seemed to strive almost exclusively for rhythm and speed. In all her drill work she counted or beat time for the pupils. It is not strange, therefore, that the pupils in this community wrote better when told to hurry than when told to use their best handwriting.

The important consideration to be emphasized is that the instructions given have a great influence on the quality of the resulting handwriting. This raises a serious objection to any effort to measure both speed and quality in the same test. The directions for rapid writing given on a previous page are typical of such combined tests. The number of letters written in the two minutes allowed may be divided by two to give an index of speed, while the quality of the product may be rated on a standard scale. But examiners may read the same directions with such different emphasis and manner as to get quite different results from the same pupils. Even the same examiner may on one occasion stress the phrase " Write as rapidly as you can" and on another occasion emphasize especially the words " Do it as well as you can." One would be more certain of his comparisons of the quality of writing in different places if quality were the only element sought. Similarly, in securing rate scores for comparison between schools, nothing but speed should be emphasized in the instructions. The writer believes that the important thing to be known is not how rapidly a pupil can write when he tries, or how well he can write when he makes special effort, but how well he writes in ordinary composition and correspondence.

Proposed objectives for teachers of elementary-grade handwriting. — On the basis of the experience that is

available now from measurements of handwriting in scores of school systems, the following tentative objectives are suggested for instruction in handwriting in the elementary grades. The scores for speed indicate the median number of letters per minute in a two-minute speed test conducted in the manner described on page 183. Care should be taken in giving the directions to see that no special emphasis is given to the phrase "as well as you can." This test

TABLE XXVI

HANDWRITING OBJECTIVES FOR THE ELEMENTARY GRADES

GRADE	II	III	IV	V	VI	VII	VIII
Speed, Best	35	45	55	64	72	77	80
Quality, Usual	7.0	7.8	8.6	9.3	9.9	10.5	11.0

should be clearly a trial of speed in writing a familiar sentence over and over again.

The scores for quality in Table XXVI indicate the median quality on the Thorndike Scale used by pupils in their English composition papers or other regular work. It is not expected that second-grade pupils will maintain a median quality of 7.0 while writing at a median speed of 35. The two objectives are to be used separately rather than in combination. Speed is to be measured by one test and quality by another. If a teacher gives her class the opportunity to write a few lines of their best handwriting, the median quality of the resulting samples should be at least one step higher than the objectives for usual quality given in Table XXVI.

There are many schools in which higher writing attainments have been secured than those proposed in Table XXVI. Handwriting is a subject in which effective drill can quickly bring about substantial improvements. The objectives proposed are merely the reasonable expectations that one may hold for each grade at about the middle of the second half of the year. It is anticipated that half of the pupils in a grade will not be able to reach the quality and speed suggested, except by an amount of drill and effort which might be considered excessive. The medians proposed are conservative estimates of what the average American school system may reasonably achieve in each grade without giving an undue amount of time and attention to this subject.

Calculation of percentile points in handwriting tabulations. — When each pupil's paper has been rated as to quality, frequency tables may be prepared showing how many pupils of each age in each grade have made each possible score. Such a table would contain one distribution of scores made by nine-year-old pupils in the fourth grade, another of the scores made by ten-year-old pupils in the fourth grade, and so on for each age found in the fourth grade, with a summary distribution showing the frequency of each size of score for all ages in this grade combined. There would be similar distributions by ages in the fifth grade, with a summary of all fifth-grade scores, and similarly for each grade in which the test was given. The same distributions should then be summarized by ages, including in each age distribution the scores of all pupils of that age in any and all grades.

As an example of the frequency distributions from which

further calculations would be made, one may take the handwriting-quality scores obtained by the twelve-year-old pupils in the sixth grade at Baltimore. This distribution is provided by Table XXVII.

To show the statistical calculations with handwriting scores, percentile points may be worked out from Table XXVII. The ten percentile or first decile point may be considered first. One-tenth of 486 scores would be 48.6 scores. To include the lowest 48.6 scores would mean taking all of those having a score of 7, all of the 32 having a score of 8, and 9.6 of the 115 who have a score of 9 (7+32+9.6 = 48.6). Now 9.6 cases are .075 of the 115 "quality 9" cases. It will therefore be necessary to pass over .075

TABLE XXVII

HANDWRITING SCORES (THORNDIKE)
TWELVE-YEAR-OLD PUPILS IN BALTIMORE SIXTH-GRADE CLASSES

SCORE	FREQUENCY
14	1
13	50
12	67
11	98
10	116
9	115
8	32
7	7
6	0
Total	486

of the quality range called 9 in order to reach the ten percentile point. The lowest limit of the quality called 9 was

demonstrated on page 180 (Figure 28) to be 8.5. By the time one has gone .075 of a step beyond 8.5, he will have arrived at a value of 8.58 (8.5 + .075 = 8.575), which is the ten percentile point.

It will be observed that the calculation of the ten percentile point from this distribution of handwriting scores is exactly like the calculation of a percentile point from a distribution of spelling scores (see page 156), except that a score of 9 in handwriting means from 8.5 up to 9.5, while a score of 9 in spelling means from 9.0 up to 10.0. This difference is due to the dissimilarity of method in obtaining scores. In rating handwriting quality, a written sample is assigned the numerical value of the sample on the scale which most nearly equals it in merit; while in rating spelling ability a pupil is assigned a score which indicates how many words are spelled with perfect accuracy, no allowance being made for a word that is partly right. This distinction corresponds to that between counting ages to the nearest birthday and counting them to the last birthday (see page 90).

The fifty percentile or median point is, of course, calculated in a manner similar to that described above for calculating the ten percentile point. One-half of 486 is 243. To include the lowest 243 of the 486 scores, one would need to take all those with scores of 7, 8, and 9 (7 + 32 + 115 = 154), and 89 of the 116 with scores of 10 (243 − 154 = 89). Now 89 cases are .767 of the 116 "quality 10" cases, so in order to reach the fifty percentile point one will need to pass .767 of the distance from the low end of quality 10 to the high end. Since the lowest possible value of 10 is 9.5, and since the point sought is .767 of a step above this, 10.3 is the median point (9.5 + .767 = 10.267).

Calculation of mean values in handwriting tabulations. — The method of assigning scores to individual papers has an effect upon the calculation of the mean point in a distribution similar to that it has on the determination of percentile points. In calculating the mean score from a distribution of spelling scores, the mean or middle value of each score was first multiplied by its frequency. The

TABLE XXVIII

CALCULATION OF A MEAN HANDWRITING VALUE

SCORE	FREQUENCY	PRODUCT
14	1	14
13	50	650
12	67	804
11	98	1078
10	116	1160
9	115	1035
8	32	256
7	7	49
Sum	486	5046

5046 ÷ 486 = 10.387 = Mean

same procedure is to be followed in this case. In handwriting, however, a score of 7 runs from 6.5 up to 7.5, and the mean value of the step would be 7.0, while in spelling the middle value of a score of 7 would be 7.5. The mean of handwriting tabulations therefore involves less labor in its calculation than the mean of spelling tabulations.

If the tabulation in Table XXVIII were of spelling scores rather than of handwriting scores, .5 of a step would need to be added to the 10.39 in order to obtain the mean

point in the distribution. Since handwriting scores are obtained by rating each child's handwriting " to the nearest value," one may without further calculation speak of 10.4 as the mean quality.

Calculation of range of middle fifty per cent. — Although the method by which scores for handwriting quality are assigned makes the location of the mean and percentile points .5 of a score lower than they would be if the distributions were of spelling scores, it makes no change in any measure of dispersion. In Table XXVII, for example, the limits of the middle fifty per cent are 9.22 and 11.46. If Table XXVII had presented spelling rather than handwriting scores, the limits of the middle fifty per cent would have been 9.72 and 11.96. In either case, however, the range of the middle fifty per cent remains the same, 2.24 (11.46 − 9.22 = 11.96 − 9.72 = 2.24). If one were employing the range of the middle twenty per cent as a measure of dispersion, his result for Table XXVII would be .87, whether the limits of this middle twenty per cent were 10.35 and 11.22, as in spelling scores, or 9.85 and 10.72, as in handwriting scores. One may bear in mind as a helpful generalization that the particular method used for grouping one's measures affects their average and other point scores but does not change their measures of dispersion or variability.

Other scales for measuring the quality of handwriting. — About two years after the appearance of the Thorndike Scale for rating the quality of handwritings, another scale of similar purpose appeared. In many respects the Ayres Measuring Scale for Handwriting[1] was an improvement

[1] Published by the Russell Sage Foundation.

upon the Thorndike Scale. It was easier for the average teacher to see the differences between the steps on the Ayres Scale, for these differences were larger than the differences on the Thorndike Scale. The Ayres Scale represented only eight different degrees of quality in handwriting, while the Thorndike Scale had represented fourteen different degrees. The Ayres Scale contained three styles of writing at each step, while the Thorndike Scale, although it contained samples of both vertical and slant writing, did not have samples of these different types for each step on the scale. The Ayres Scale was printed in blue ink which gave it the appearance of having been done with a pen, while the Thorndike Scale was printed in black ink throughout.

It is also probable that the numerical values connected with the Ayres Scale samples, being 20, 30, 40, 50, 60, 70, 80, and 90, increased the popularity of this scale more rapidly than it would have increased had the values been wholly lacking in their suggestion of a percentage scale. Teachers seem to enjoy assigning a mark of 80 more than they enjoy assigning a mark of 7 or of 15, even though the meaning of the briefer mark may be exactly the same as that of the larger. Another factor that made the Ayres Scale popular was its more convenient size and shape. The Ayres Scale is on a strip of paper about 3 feet long but only 9 inches wide, which fact makes it relatively easy to handle.

The so-called Gettysburg Edition of the Ayres Scale, first published in March, 1917, differs from the original three-slant edition in that it contains only one style of writing, a semi-slant type. The name of the later edition is due to the fact that each of the eight steps is a copy of the

first few sentences of Lincoln's Gettysburg address. The Gettysburg Edition carries with it printed directions for securing samples of handwriting, for measuring the speed at which these are produced, and for measuring their quality. Standard scores and distributions are also provided. For introducing teachers who have never used a scale to the method of rating the quality of handwriting, there is no more satisfactory scale than the Gettysburg Edition of the Ayres Scale.

The Starch Measuring Scale for Handwriting [1] has most of the advantages of the Gettysburg Edition of the Ayres Scale, but it is probably too expensive to meet with extended classroom use. One cause of its high price is the fact that it is printed on cardboard rather than on paper. The Starch Scale represents twenty different degrees of quality in handwriting. Both capital letters and small letters appear at each step from 0 to 20. For the use of trained persons in experimental work the Starch Scale is possibly somewhat more satisfactory than the Thorndike Scale, but it will never be as popular as the Ayres Scale or the Thorndike Scale in general school work.

All of the scales so far mentioned are intended to measure the general merit of handwritings. Professor Frank N. Freeman has prepared a Chart for Diagnosing Faults in Handwriting,[2] which affords the teacher a means of analyz-

[1] Copies of this scale may be obtained from the University Cooperative Company, 504 State Street, Madison, Wis. The derivation of this scale is described by Daniel Starch, in "A Scale for Measuring Handwriting," *School and Society*, January, 1919.

[2] For further discussion, information, and directions, see Frank N. Freeman, *The Teaching of Handwriting*.

ing the elements that cause a given sample of writing to be poor or good. The Freeman Chart is essentially a set of five handwriting scales, in each of which there are three degrees of merit. The lowest quality in each is numbered 1, and the highest quality is numbered 5. The medium quality is numbered 3. If one desires to do so, he may rate a given sample of writing as 4 when it is better than quality 3 but poorer than quality 5. Similarly a rating of 2 may be given when 3 would be too high and 1 would be too low.

The first scale at the top of the Freeman Chart is intended to measure the uniformity of the slant in a sample of handwriting. The second scale measures uniformity of alignment. The third is planned to show differences in the quality of line used. The fourth scale is for the formation of the letters, and the fifth is for spacing.

It is suggested by Freeman that one may obtain a total score for an individual sample of handwriting by adding together the separate scores obtained on the five scales, but such a pooling of scores seems to be something of a waste of valuable information. If a general score is wanted for a page of handwriting, it may be secured by comparing it with the Thorndike, the Ayres, or the Starch Scale. The important function of the Freeman Chart is to point out the particular respects in which a given pupil's handwriting can be improved.

There have been numerous other handwriting scales or local supplements to scales described above, but the Thorndike and the Ayres have been used most extensively. The chief defect of the Thorndike Scale in its original form was an inadequate representation of quality 10. Many school

surveys and other reports containing tabulations of Thorndike scores show a much smaller frequency of 10's than of 11's and 9's, merely because the examiners did not find at 10 on the scale an example having the popular slant. This defect has been corrected in the scale as now issued. The Thorndike Scale provides for somewhat finer distinctions in quality than the Ayres. It is somewhat easier, however, to teach untrained persons to rate handwriting by means of the Gettysburg Edition of the Ayres Scale than by the Thorndike Scale.

Summary. — Handwriting under the conditions of modern life has lost much of its standing as a fine art and has become a more or less prosaic tool for social and industrial communication. The Thorndike Scale was developed to measure the general merit of children's handwritings. Directions are supplied in this chapter for obtaining a sample of the usual handwriting of each pupil, and for assigning scores for quality in terms of the Thorndike Scale. The handwriting results obtained in a number of city school systems are reported by age and by grade, and tentative objectives for the various elementary grades are proposed. Brief comments are made on the Ayres and on the Starch Scales, and on the Freeman Chart for Diagnosing Faults in Handwriting.

QUESTIONS AND EXERCISES

1. What influence has the perfection of the typewriter probably had upon objectives in the teaching of handwriting?

2. What reasons can you suggest for not training pupils to write rapidly until after they have learned to write legibly?

3. Why should a special exercise, such as the composition assign-

ment on page 173, rather than some ordinary examination paper in history or geography be used as a source of handwriting samples?

4. What conditions might cause the 10th, 11th, and 12th lines of a composition to be unfair samples of a pupil's handwriting?

5. Is it your understanding that quality 10 on the Thorndike Scale is twice as good as quality 5? Would it take more or would it take less instruction to improve the score of a pupil from 5 to 10 than from 7 to 12? Why?

6. Why should the final score on a sample of handwriting be based on the independent ratings of three judges rather than on three judgments by the same person?

7. Twenty pupils' handwritings were rated on the Thorndike Scale, as follows: 10, 9, 11, 10, 11, 8, 10, 11, 10, 12, 9, 10, 10, 9, 11, 10, 8, 11, 12, 10. Calculate the median and other decile points. What is the range of the middle fifty per cent? middle 80 per cent?

8. Test your pupils for quality of usual handwriting, quality of best handwriting, and quality of most rapid writing, according to the directions on pages 173 and 183. What do the differences in median scores obtained under these three conditions indicate with reference to the training your pupils have had in handwriting?

9. In rating handwritings to the nearest sample on the Thorndike Scale, all samples between 10.5 and 11.5 would be called 11, all between 11.5 and 12.5 would be called 12, and so on. If samples were similarly assigned to the nearest scale value on the Ayres Scale, what range of values would be called 40? What would be the limits of a score of 50?

REFERENCES

L. P. Ayres, *A Scale for Measuring the Quality of Handwriting of School Children*, page 16.

F. N. Freeman, *The Teaching of Handwriting.*

T. L. Kelley, *Statistical Method*, pp. 115–119.

D. Starch, "A Scale for Measuring Handwriting," *School and Society*, Vol. XI, pp. 154–158 and 184–188.

E. L. Thorndike, "Handwriting," *Teachers College Record*, Vol. II, No. 2.

CHAPTER IX

AVERAGE MEASURES

NOT all members of a group or class make exactly the same score on a standard test, but most of them obtain a result that is not far from the average score. To compare two groups with each other usually requires a comparison of their total distributions of scores. It is sometimes convenient, however, and for many purposes entirely sufficient, to compare the average scores made by the two groups. Still more frequently, all practical purposes may be served by comparing first the measures of average and then the measures of compactness or dispersion shown by the two groups. Two classes might be exactly equal in their average scores, yet one class might be more satisfactory than the other by reason of having fewer extremely high and fewer extremely low scores. A scientific writer seldom reports an average score without accompanying it by a measure of the dispersion of the group measured.

This chapter undertakes to review and discuss the important considerations regarding average measures, and the two following chapters attempt to show how variability or dispersion may be determined and represented. Although presented here in different chapters, measures of central tendency and measures of dispersion are different characteristics of the same groups of scores.

Classification of average measures. — Measures of central tendency or average may be reported with great mathematical precision, or they may be stated in general terms. One may calculate the exact *point* on the scale which represents the central value for a group of measures, or he may indicate only the *score*, or possibly only the *class* of scores, in which this central value would be found. Whether the exact point or only the approximate location is desired depends, of course, on the nature of the original measures and upon the purpose for which the central tendency is being determined. One should always be conscious, however, of the degree of accuracy he is seeking. Above all, one should not use a formula intended only for locating a central class of scores as though it were a rule for locating an exact central value. Needless confusion and error have resulted from overlooking this distinction.

Three measures of central tendency or average will be considered: mode, mean, and median. The *mode* is merely the most frequent measure or size of score. It is (*la mode*) the typical thing, the fashion, the measure which is most in vogue. The *mean*, often known as the arithmetic average, is the quotient obtained by dividing the sum of all the measures by their number. The size of each individual measure in the group therefore influences the size of the mean. The *median* is the middle size in the group when the measures are arranged according to size. It is the value above and below which there are equal numbers of measures. It is unlike the mean in that it is not influenced by the actual sizes of all the measures in the group but is influenced by their number and by the sizes of the middle measures.

Scores for speed in handwriting. — A fifth-grade boy was tested for two minutes each day, except Sunday, over a period of several weeks. The test applied was that described on page 183 of the previous chapter. Table XXIX gives the number of letters written each day, fractional parts of letters having been carefully estimated by the examiner as to the degree of their completeness. The *score* for each test, tabulated in Table XXX, is half the number of letters written in two minutes. Any letter that was incomplete at the end of the two-minute period was not counted in the scoring. Where an odd number of letters was written in the two-minute period, resulting in a score for one minute that included a half letter, the fraction was dropped before the score was recorded. The

TABLE XXIX

DAILY RECORD OF HANDWRITING SPEED TEST RESULTS

DAY	LETTERS IN 2 MINUTES	SCORE	DAY	LETTERS IN 2 MINUTES	SCORE	DAY	LETTERS IN 2 MINUTES	SCORE
1	118.8	59	13	123.1	61	25	129.0	64
2	123.7	61	14	127.8	63	26	131.6	65
3	126.1	63	15	132.0	66	27	136.6	68
4	124.2	62	16	131.5	65	28	133.4	66
5	121.1	60	17	128.8	64	29	129.3	64
6	122.4	61	18	129.7	64	30	126.8	63
7	117.5	58	19	126.7	63	31	129.9	64
8	121.9	60	20	135.1	67	32	138.7	69
9	128.4	64	21	132.4	66	33	133.6	66
10	127.2	63	22	135.4	67	34	134.5	67
11	130.9	65	23	130.3	65	35	128.2	64
12	125.8	62	24	124.0	62	36	131.0	65
Sum	1488.0	738	Sum	1556.8	773	Sum	1582.6	785

TABLE XXX

FREQUENCY DISTRIBUTION OF HANDWRITING SPEED SCORES

SIZE OF SCORE	FREQUENCY OF SCORE	FREQUENCY FROM BEGINNING
69	1	36
68	1	35
67	3	34
66	4	31
65	5	27
64	7	22
63	5	15
62	3	10
61	3	7
60	2	4
59	1	2
58	1	1
Total	36	

tabulated score for each day was therefore the mean number of letters written in one minute, all fractional parts of letters having been dropped.[1]

Section A, the upper part of Figure 29, represents the daily records of the boy's handwriting speed, in terms of

[1] It would have been just as reasonable to have tabulated each fraction of a letter as though it were a whole letter; or to have considered each less-than-half-finished letter as non-existent and each more-than-half-finished letter as a completed letter. It is easier for the average teacher, however, to throw aside partially completed letters than it is to decide when a child's mark was intended for the beginnings of a letter or to determine just what fraction of a complete letter is represented by a given fragment. It is necessary that a definite rule be adopted for tabulating these fractional results and that the rule be considered when calculating statistical points from the tabulations.

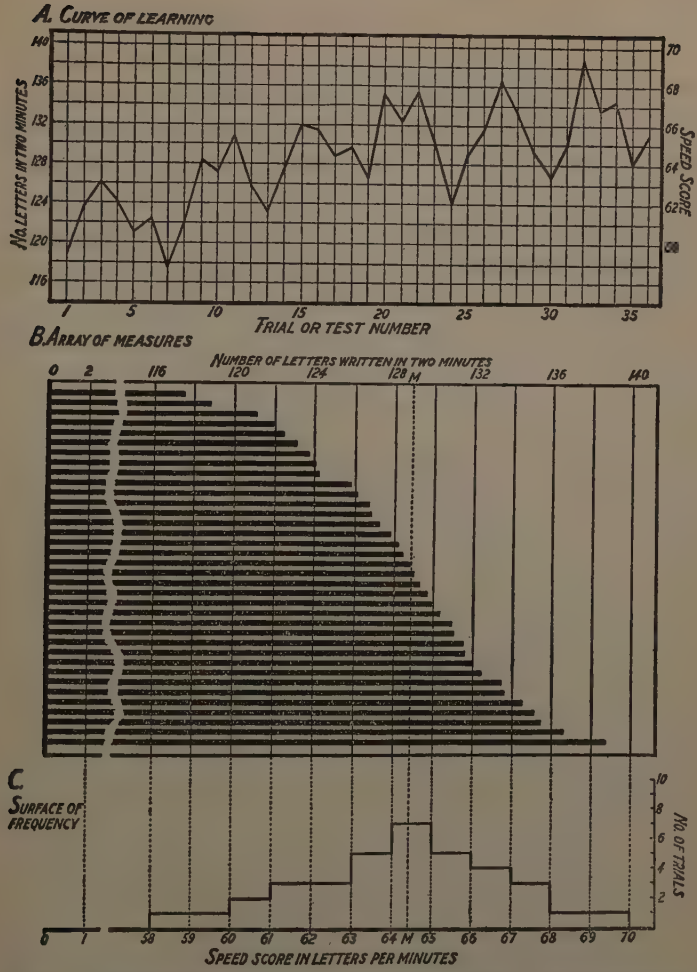


FIG. 29. — A fifth-grade boy's speed in handwriting. Section A shows the changes in speed from day to day. Section B shows the different speed records arranged in order of size. Section C shows a distribution of the scores, grouped to the last full letter per minute. Dotted line MM, erected at the median point in the distribution, divides the surface of frequency into two equal areas and the array into two groups of eighteen measures each.

letters (and fractions of letters) formed in two minutes. The height of the curve on any particular day represents the number of letters formed that day, as recorded in Table XXIX. It will be noted that there is a general tendency for the later records to be higher than the earlier ones, although the progress made was very irregular. There are distinct peaks and depressions in the curve, probably representing a weekly rise and fall in ambition, energy, or ability. The general tendency toward higher scores undoubtedly represents improvements in handwriting speed due to the practice obtained in the earlier tests.

Section B, the middle part of Figure 29, represents the same measurements as were shown in section A. The number of letters per minute are represented by the lengths of the horizontal black bars in section B, while section A represents the same facts by the perpendicular heights of a broken curve. The most significant difference between these two sections of Figure 29, however, is that in section A the measurements are given in the order in which they occurred, while in section B they are " arrayed " in the order of their magnitude, from low to high speed.

Section C, the lower part of Figure 29, represents a grouping into a surface of frequency of the measures shown in section B. Horizontal distance to the right represents in section C the number of letters per minute, on the same scale as in section B. The height of the curve above any particular unit of the base line in section C represents the number of days on which that particular writing speed was attained. Reading from the left end of the curve, a speed between 58.0 and 59.0 was obtained once, a speed between 59.0 and 60.0 once, a speed of 60.0 to 61.0 twice,

61.0 to 62.0 three times, and so on. Faint lines have been drawn to indicate the grouping in section C of the individual measures represented in section B.

Section C represents graphically the table of frequencies given in Table XXX, in which a score of 69 means from 69.0 up to 70.0, a score of 68 means from 68.0 up to 69.0, and so on. The reader will recognize that this grouping of scores " to the last completed letter " is due to the rules for assigning the speed scores. Had these directions assigned scores " to the nearest whole number of letters," the faint lines from section B to section C would have been half a letter's distance to the left of their present locations, and a score of 69 in Table XXX would have meant from 68.5 up to 69.5 letters. The method of grouping measures does not affect the actual sizes of the measures in the least, but it sometimes has a distinct effect upon the frequency surface and distribution. One must use care, therefore, in making calculations from distributions of scores, to see that the character and meaning of the original measures are not forgotten or misrepresented.

Determining the mode. — The mode is the size or measure which occurs most frequently. It is seldom determined with great precision, for the location of the exact point on the scale which best represents the mode is the most complex and difficult of all the calculations of central tendency.[1] The location of the modal group, how-

[1] The exact determination of the point on the scale which would best represent the size of the mode involves fitting a theoretically smoothed frequency curve to the one actually obtained. This process is too complex, and the resulting mode is too seldom of importance to warrant including it here. For practical purposes one may find the

ever, is easier than any other determination of averages, for it may be done by a casual inspection of the table or surface of frequency. In the second column of Table XXX, for example, 7 is the largest frequency recorded and occurs opposite a score of 64. The modal score for this distribution of speed scores in handwriting is therefore 64 letters per minute.

The same fact exactly may be observed by a glance at section C of Figure 29, which is a graphic picture of Table XXX. The highest point in the frequency curve of this section gives the location of the mode. The peak of the curve is above 64, indicating that 64 letters per minute is the mode for the distribution represented by this figure.

As was indicated in a previous chapter, tabulations are frequently made with steps or classifications somewhat larger than the scores on the original scale. Table XXXI is a tabulation by larger groups of the spelling scores of the same 15-year-old Baltimore pupils whose scores were tabulated in greater detail in Table XIX, page 159. The distribution in Table XIX shows that the crude mode for these pupils was 16 words correctly spelled. From the distribution in Table XXXI, it is impossible to locate the mode any more exactly than that it is in the step called 16, which means that it is somewhere between 16.0 and 17.99 words spelled. Except by the elaborate statistical methods

difference between the median and mean points of the distribution, and then measure off on the opposite side of the median point twice that difference. In most of the distributions in which all three averages are not practically equal, the exact mode is on the opposite side of the median and twice as far from it as the mean.

TABLE XXXI

SPELLING SCORES OF FIFTEEN-YEAR-OLD PUPILS

STEP	FREQUENCY	LIMITS OF STEP
24	3	24.0 —
22	11	22.0 — 24.0
20	46	20.0 — 22.0
18	57	18.0 — 20.0
16	100	16.0 — 18.0
14	61	14.0 — 16.0
12	41	12.0 — 14.0
10	22	10.0 — 12.0
8	17	8.0 — 10.0
6	9	6.0 — 8.0
4	2	4.0 — 6.0

used for finding the exact mode, it is not expedient to attempt to locate the mode more accurately than to be able to say that it is in a certain step or group. The coarser grouping is sometimes helpful in locating a crude mode, however, especially when finer grouping shows very few cases or frequencies on each step.

Bimodal distributions. — Occasionally one finds that a distribution of measures has more than one mode or peak. In such a case it is spoken of as a multimodal distribution. The most common type of multimodal distribution is the bimodal, which has two modes. The usual cause for a bimodal distribution is that two separate distributions have been combined, either intentionally or accidentally, as in the total column of Table XXXII. This table presents the distributions of handwriting scores found in the fifth grades of two neighboring towns. The

mode at P. W. was quality 9 and the mode at V. S. was quality 11 on the Thorndike Scale. When a summary of the scores of both schools was made, there was one mode at 9 and another at 11.

A bimodal distribution does not always indicate that two distinct groups have been combined. The bimodal distribution in Table XXXII might have resulted from the

TABLE XXXII

FIFTH-GRADE HANDWRITING SCORES

QUALITY (THORNDIKE)	FREQUENCY		TOTAL
	At P. W.	At V. S.	
13	—	2	2
12	—	9	9
11	1	20	21
10	12	4	16
9	20	1	21
8	8	—	8
7	1	—	1

action of two powerful but independent factors, each acting upon the same group of pupils. The same distribution might, on the other hand, have been the result of a careless rating of handwritings on the older edition of the Thorndike Scale, in which quality 10 was of such a character as to be seldom used by untrained judges. Whenever a distribution is clearly bimodal, it is wise to try to discover the causes, being especially on the alert for the possibility that two or more distinct groups have been combined.

Calculating the mean value of untabulated measures. — The mean, or the average as it is more frequently called

in America, is the sum of all the values divided by their number. The problems in calculating averages presented by the usual school arithmetics are not difficult to solve, for each numeral used in such a problem can be taken at its face value. Ten means exactly 10.0, no more and no less. The average class, in a school of five classes having in all 100 pupils, contains 20 pupils ($100 \div 5 = 20$). The average daily wage of three men whose combined income is $12 per day can be quickly calculated as $4 per day ($12 \div 3 = 4$). If one train runs 30 miles per hour, another runs 32 miles per hour, and a third 25 miles per hour, their mean rate is asserted to be 29 miles per hour ($\frac{1}{3}$ of 87 $= 29$).

The calculation of mean scores in standard educational tests is not essentially different from the simple calculations given above. One must bear in mind, however, that educational scores are not usually reported with great precision but are so grouped that a score of 5 does not mean exactly 5.0 but means somewhere between 5.0 and 6.0; or possibly from 4.41 to 5.48, as it does in one English composition scale. As a matter of fact, the statistician who desires to make accurate comparisons of the rate of a certain train with the average rate of the three trains mentioned in the above paragraph, would ask for the exact definitions of rate by which the average was found to be 29 miles per hour. He would want to know just what 29 means. He would, therefore, seek to learn by what rules the three original rates were determined. Was a train rated at 30 miles per hour if its rate were nearer 30.0 than to 29.0 miles per hour, or did 30 miles per hour mean more than 30.0 but less than 31.0 miles? If the

three original rates were stated *to the nearest mile*, then the mean of the three values is 29.0; but if the original scores were stated *to the last full mile*, the arithmetic average of their values is 29.5 miles $\left(\frac{30.5 + 32.5 + 25.5}{3} = 29.5\right)$.

With the same general considerations in mind, one may seek the mean of the speed scores for the fifth-grade boy in Table XXIX. The boy's score on the second day was 61, although his actual speed was about 61.8 letters per minute. On the third day his score was 63, while his actual speed was 63.05 letters per minute. A score of 60 includes, by the conditions of the scoring, everything from 60.0 up to 61.0 letters per minute, and a score of 61 includes from 61.0 up to 62.0 letters per minute. A score of 60 would therefore tend to include as many records below 60.5 as above 60.5 letters per minute. Similarly, the best value to represent a score of 61, having a range from 61.0 to 62.0, would be 61.5 letters per minute. To add the scores of Table XXIX would not, therefore, be adding their most probable mean values. The mean values of the thirty-six scores would be .5 of a letter higher than their face values in every case. In other words, the sum of the thirty-six face values divided by thirty-six would give a quotient that was .5 of a letter too small to represent the sum of the mean values divided by their number.

Knowing that calculations with the face values would give a result exactly .5 too low, one need not necessarily take the trouble to add .5 to each separate score in Table XXIX in order to obtain the true mean. It is much more economical to use the scores at their face value and then to add .5 to the quotient finally obtained. Table XXIX happens to be in three sections, so arranged that one may

conveniently add the scores of the first two weeks separately from those of the second two weeks and from those of the final two weeks. The sum of the scores for the first period is 738, for the second period 773, and for the third period 785. Each period contained the scores from 12 tests. The quotients obtained by dividing the sums by 12 are 61.5, 64.4, and 65.4 respectively. Adding .5 to each of these quotients, in order to correct for having used the face values rather than the mean values, shows that the mean speed during the first period was 62.0, during the second period 64.9, and during the final period 65.9 letters per minute. The mean speed score for the entire series of thirty-six trials, found by adding .5 to the quotient obtained by dividing 2296 (the sum of 738, 773, and 785) by 36, is 64.28 letters per minute.

This mean rate, obtained from the scores of Table XXIX, may be compared with the mean rate that would be obtained by using the more exact measurements of the number of letters written in two minutes. The total number of letters written, obtained by adding the thirty-six records in Table XXIX, was 4627 letters. Dividing this sum by seventy-two, the total number of minutes, gives 64.27 letters per minute as the exact mean rate of the boy's writing. As far as the mean is concerned, therefore, one obtains in this case almost exactly the same result from calculating with the scores in round numbers as from calculating with the larger numbers representing exact values, provided the scores have been assigned by definite rules and then treated carefully in the light of these rules.

Calculating the mean value of tabulated data. — One of the advantages claimed for the mean over other measures

of average is that the mean can be easily secured without throwing the data into an array or table of frequency. This possibility, however, of securing an average measure from wholly unorganized records is sometimes a disadvantage. Inexperienced persons frequently obtain an average score in this manner, without a clear conception of the total distribution of scores, and then too often they think of the mean value thus obtained as typical of the entire situation. In some cases the mean value may even give a false impression of the character of a variable measure. If time permits, therefore, and especially if the data are important, records of measurements should be tabulated to show the frequency of the scores of various sizes, before even the mean is calculated. This will tend to keep one from forming false conclusions from his average results.

The calculation of a mean value from a frequency table involves only one point that is not involved in its calculation from the untabulated data. Table XXXIII shows the steps involved in the calculation of the mean speed of handwriting from the scores given in Tables XXIX and XXX. Instead of adding the 36 individual scores separately, as was done above (page 209), the three scores of 67 are listed in column III as 201 ($3 \times 67 = 201$), the four scores of 68 are listed as 264, and so on, the size of the score being multiplied in each case by its frequency. The sum of these products is, of course, the same as the sum that was obtained by adding the individual scores separately, 2296. From this point forward the calculation is exactly the same as was described on page 209.

Table XXXIV shows the calculation of the mean quality of handwriting in the fifth grades of each of the two villages

TABLE XXXIII

Calculation of Mean Speed Score in Handwriting

I Score	II Frequency	III Product
69	1	69
68	1	68
67	3	201
66	4	264
65	5	325
64	7	448
63	5	315
62	3	186
61	3	183
60	2	120
59	1	59
58	1	58
Sum	36	2296

$2296 \div 36 = 63.78$; $63.78 + .5 = 64.28$

TABLE XXXIV

Calculation of Mean Fifth-Grade Handwriting Quality

Quality (Thorndike)	Village P. W.		Village V. S.	
	Frequency	Product	Frequency	Product
13	—	—	2	26
12	—	—	9	108
11	1	11	20	220
10	12	120	4	40
9	20	180	1	9
8	8	64	—	—
7	1	7	—	—
Sum	42	382	36	403
Mean	$382 \div 42 = 9.1$		$403 \div 36 = 11.2$	

represented by the scores in Table XXXII. Since the scores in handwriting were assigned to the nearest quality on the scale, a score of 10 means from 9.5 up to 10.5 and has its central value at 10.0. The face values of the scores in handwriting are the mean values in each case and are therefore to be multiplied by their frequencies in calculating the mean quality in the distributions. In other words, the sum of the products of face values and frequencies, divided by the number of scores, gives at once the mean handwriting score one is seeking.

Where scores have been tabulated into groups larger than the units of the original scale, the process of finding the mean point is not essentially different from that outlined above. One must use care, however, to correct for the difference between the face value and the central value of each class or step in the tabulation. When a classification group extends from 24.0 up to 26.0, its central value is 25.0. If 24 has been used as the face value with which to multiply its frequency, then one must, in order to find the mean, add 1.0 as a correction to the first quotient. When a step in the tabulation extends from 20.0 up to 25.0, its central value is 22.5. If 20 has been used as the face value in multiplying, a correction of 2.5 must be added, while if 25 has been used in multiplying, a correction of — 2.5 must be added to the first quotient in order to find the mean.

Table XXXV (*a*) shows the calculation of the mean for the spelling scores of fifteen-year-old pupils in Baltimore. These scores are the same that were given in Table XXXI and in Table XIX, although in this tabulation the steps are coarser, each containing a range of four units of the original

list of 25 words. The central value of each group or step in the tabulation is 2.0 words above the initial value used in the multiplications, so that 2.0 must be added to the first quotient in order to locate the mean spelling ability.

TABLE XXXV (*a*)

CALCULATION OF MEAN SPELLING SCORE OF FIFTEEN-YEAR-OLDS

RANGE	STEP	FREQUENCY	PRODUCT
22–25	22	14	308
18–21	18	103	1854
14–17	14	161	2254
10–13	10	63	630
6– 9	6	26	156
2– 5	2	2	4
Sum		369	5206

5206 ÷ 369 = 14.1; 14.1 + 2.0 = 16.1 = Mean

In order to avoid dealing in large numbers, the steps of the tabulation might be numbered 6, 5, 4, 3, 2, and 1, as in Table XXXV (*b*), rather than 22, 18, 14, 10, 6, and 2. In this case the multiplications would be easier, but one would need to keep in mind that the first quotient (4.03) was in terms of the new names of the steps, and that 4.0 in these terms stands for the central value (16.0) of the range 14.0 up to 18.0. The quotient 4.03 then means "at .03 of a step above the central value of step 4," which would be at .1 of a word above 16.0 words. This scheme of reducing the size of the multiplications by substituting small numbers for the original scores is of great assistance

where the original scores are too large for easy mental multiplications. The danger is that beginners will neglect to convert their results back into terms of the original scores.

TABLE XXXV (*b*)

Step	Frequency	Product
6	14	84
5	103	515
4	161	644
3	63	189
2	26	52
1	2	2
Sum	369	1486

1486 ÷ 369 = 4.03 (steps); Central value of step 4 = 16.0 words; .03 step = (.03 × 4 = .12) .1 word; 16.0 + .1 = 16.1 words = Mean

As another example of this substitution, Table XXXVI shows the calculation in this manner of the mean speed in handwriting from the scores previously given in Tables XXIX, XXX, and XXXIII. When the scores are in extremely large figures and are grouped into classifications each containing a fairly wide range of scores, the temporary assignment of new names to the classifications aids both the accuracy and the speed of calculating mean points in distributions.

Thus far nothing has been said in this chapter regarding mean scores. All the discussion has been regarding mean values of measures or mean points in distributions. In order to avoid confusion, it will be better for the beginning student in statistics to make all initial calculations of means

TABLE XXXVI

Calculation of Mean Speed Score in Handwriting

Score	Step	Frequency	Product
69	12	1	12
68	11	1	11
67	10	3	30
66	9	4	36
65	8	5	40
64	7	7	49
63	6	5	30
62	5	3	15
61	4	3	12
60	3	2	6
59	2	1	2
58	1	1	1
Sum		36	244

244 ÷ 36 = 6.78 (steps); Central value of step 6 = 63.5 letters .78 step = .78 letter per min.; 63.5 + .78 = 64.28 letters = Mean

or of medians in terms of points or values rather than in terms of cases, scores, or steps. The mean score is nothing more than the score containing the mean value or point. Similarly, the mean step in the distribution is the step in which the mean point is located; and the mean case in a group is the case having the mean value (if any case happens to have that exact value). In Table XXXIII, for example, the mean speed is 64.28 letters per minute. The mean score is therefore 64, for it covers the range from 64.0 up to 65.0 letters per minute. The mean step in Table XXXVI is step 7, which in this table has the same limits and the same

meaning as a score of 64. The mean step in Table XXXV (*b*) is step 4, and the mean score is 16, for the mean spelling ability, 16.1 words, falls within the range (16.0 – 17.0) to which a score of 16 is attached by the directions for marking the test.

Calculating the median value of tabulated data. — One of the arguments for using the fifty percentile or median point as an average in dealing with educational facts is that it is necessary to organize the individual measures either into an array or into a table of frequency before the median value can be conveniently calculated. When one sees the total range of the measures, as he must in an array or distribution, he is not likely to think of an average measure as representing adequately all the others. The very definition of a median value is in terms of an organization of all the values in the order of their relative sizes. The median value is the middle value when all measures are arranged in order of size.

In calculating the median value from a distribution of measures, one must make the same assumptions, regarding the different measures that are registered on a given step of the distribution, as were made in calculating the mean value.[1] If there are ten scores of 23 in spelling, one must assume that they are spread evenly over the entire range from 23.0 to 24.0.

Each median or mean value for the distributions of Table XXXVII will be .5 higher if these scores represent spelling results than it will be if they represent quality of handwriting results, for the limits of spelling steps are .5 higher (pp. 180, 190). The scores in Table XXXVII are the

[1] See Fig. 18, p. 122.

TABLE XXXVII

Distributions of Sixth-Grade Scores

Size of Score	Frequency In School							All Schools
	A	B	C	D	E	F	G	
14	—	—	—	—	1	—	1	2
13	—	—	1	3	1	1	3	9
12	—	2	1	6	2	—	2	13
11	5	4	7	—	—	4	—	20
10	7	7	1	5	2	—	—	22
9	5	—	6	2	1	2	4	20
8	—	—	2	1	1	2	2	8
7	—	—	—	—	—	—	—	—
6	—	—	—	—	—	1	—	1
Total	17	13	18	17	8	10	12	95

distributions of handwriting qualities (Thorndike Scale) found in the sixth grades of several rural schools.

School A furnishes 17 scores. To find the median quality of handwriting in this school one must count through 8.5 of them. Beginning with the smaller scores, one would have to count the 5 at quality 9 and the lowest 3.5 of the 7 at quality 10, which would take one .5 of the way (3.5 is .5 of 7) through the range of quality 10. Since quality 10 ranges from 9.5 to 10.5, one would be at 10.0 when half-way through it (9.5 + .5 = 10.0). The median quality in School A was therefore 10.0 on the Thorndike Scale.

To find the median quality in School B one must count up through 6.5 of the 13 scores. Since there are 7 scores at quality 10, one will have covered .9 of the distance from

9.5 to 10.5 by the time he counts through 6.5 scores (6.5 is .9 of 7). The median quality in School B was therefore 10.4 (9.5 + .9 = 10.4).

To find the median in School C one must count through 9 of the 18 scores. This brings one to the point that divides the score of 10 from the score of 11. Since quality 10 ranges from 9.5 to 10.5 and quality 11 from 10.5 to 11.5, the dividing point is 10.5, which was therefore the median quality in School C.

The gap at 11 in the distribution for School D does not interfere with the calculation of the median value. To count up through 8.5 of the 17 scores, one must pass all of those below 12 and .5 of one of the 6 scored at 12, which would carry one approximately .1 of the way (.5 is .1 of 6) from 11.5 to 12.5. The median quality in School D was therefore 11.6 (11.5 + .1 = 11.6).

The gap in the distribution at School E happens to come at the middle of the distribution. By the time one has counted up through 4 of the 8 scores he has reached the score of 11, but the question then arises as to the location of the median point in the range of the score called 11. The most representative point in any such range is its midpoint. The midpoint of a handwriting score of 11 is 11.0, which was therefore the median quality in School E.

The distribution for School F is to be treated like that for School E. The median falls in step 10, which was not the score of any pupil. The fact that 4 pupils had quality 11 while only 2 had quality 9 does not change the situation. The 4 pupils at 11 may be thought of as covering all the possible positions between 10.5 and 11.5, while the 2 pupils at 9 cover the range between 8.5 and 9.5. Any point

between 9.5 and 10.5 would have five scores on either side of it, but only the midpoint, 10.0, would divide the distribution (including the vacant step) into two equal parts. The median is therefore in quality 10, at its midpoint 10.0.

To divide the distribution for School G into two equal parts, one must count through 6 of the 12 scores. If one counts up from the bottom in the usual way, he has passed 6 when he reaches 9.5, the upper end of quality 9. If he counts down from the top, he has passed 6 when he reaches 11.5, the lower end of quality 12. The median point is evidently between 9.5 and 11.5. As in the cases of Schools E and F, the midpoint of the vacant range is its best representative. The median quality should therefore be recorded as 10.5, which is the division point between qualities 10 and 11.

As was previously stated, the above medians would each have been .5 of a step higher had these scores been spelling scores. The writing speed scores in Table XXX (page 200) should be treated like spelling scores, because they were recorded " to the last completed letter " rather than " to the nearest letter." By the time one has counted up through the 15 days which have scores of 63 or less, one has only 3 more to take in the next group (which contains 7 scores at 64) in order to have reached the fifty percentile or median point. By passing 3 of the 7 one will be .4 of the way ($3 \div 7 = .43$) from 64.0 up to 65.0. The median speed of this boy's writing tests was therefore 64.4 words per minute.

The median may be calculated from an array as well as from a distribution. The central drawing in Figure 29 represents in a graphic array the boy's speed measures for

36 days, including the seven measures between 64.0 and 65.0 letters per minute. The median in an array of 36 measures would be midway between the 18th and 19th measures, which in this case are the 3d and 4th scores of 64. The 18th record in order of size was made (see Table XXIX) on the boy's 17th day of practice and was approximately 64.4, while his 19th record in order of size was on the 25th day of practice and was 64.5. A speed midway between 64.4 and 64.5 would best represent the median measure of the array. A dotted line *MM* has been drawn in Figure 29 to show the relation of the median length of line in the array to the median point in a surface of frequency. Although it would be best to develop the habit of throwing all measures into frequency distributions, from which average measures and measures of variability may be calculated, it is well to note that one may easily locate a median measure from an array of scores.

The intelligent student would probably find no difficulty in calculating the median value from distributions in which each class or step was larger than a single score, but it may be worth while to give at least one example here. A score of 50 on the Ayres Handwriting Scale means from 45.0 to 55.0, because written samples are rated " to the nearest sample on the scale," as is customary in the use of all such scales for quality of products. To divide the distribution given by Table XXXVIII into two equal parts, one would have to pass the three scores at 30, the eight at 40, and seven of the twelve at 50. This would mean covering $\frac{7}{12}$ or .58 of the total range of step 50, which is 10 points long. Since .58 of 10 points are 5.8 points, the fifty percentile or median quality would be 50.8 (45.0 + 5.8 = 50.8). An essential

point to be observed in calculating the median as well as the mean value in a distribution having classes or steps each of which contains more than a single numerical score, is to remember to convert the fractional part of the step into terms of the scale scores.

TABLE XXXVIII

A DISTRIBUTION OF SCORES FOR QUALITY OF HANDWRITING (AYRES SCALE)

QUALITY	30	40	50	60	70	TOTAL
Range of Step	25.0–35.0	35.0–45.0	45.0–55.0	55.0–65.0	65.0–75.0	
Frequency	3	8	12	9	4	36

The discussion thus far has been concerned with median values or points. The median score in any distribution is the score in the range of which the median value lies, just as the mean score is the score containing the mean value. The median class or step in a distribution is the step containing the median value. In School A of Table XXXVII, for example, the median score and step are 10, because the median value 10.0 falls within the range of a score of ten. The median case or pupil is the fourth pupil who made a score of ten. In School C there is no median score, for the median quality falls just at the dividing point (10.5) between a score of ten and a score of eleven. There is no median case in School C, because there is an even number of pupils. In Table XXXVIII there is no median case, but the median step is 50, and the median score is 50, for this step and this score contain the median value, 50.8.

Formulæ for calculating averages. — The majority of those who use elementary statistical methods are not mathematically inclined and do not find it especially easy to learn operations from their formulæ. The writer has therefore avoided thus far the use of formulæ. Considering the detailed discussions in previous chapters and the review in this chapter, the assumption will be made from this point forward that the student understands and can find a mean value or a median value for any group of measures. In order to abbreviate future discussions, it is desirable that the student learn the meaning of certain notations that will be used.

$M = \frac{\Sigma\, m's}{n}$ is the formula for the mean value. The capital M is used here and will be used occasionally in the remainder of this volume as the sign of the mean value or arithmetic average. The small letter m is only an abbreviation for *measures* or score values. The small letter n stands for the number of measures or scores in the group being considered. The Greek capital letter Σ (sigma) indicates *the sum of* whatever follows it.

The above formula therefore reads as follows: "The mean value is equal to the sum of the measures divided by their number." This is, of course, exactly the way the previously given instructions tell one to find an arithmetic average: Add all the measures together and divide by their number. One needs only to remember that the name of a step in a distribution is not always its central or mean value, and that adding together the names of the scores made by pupils on an educational test may be somewhat different from adding the values of these scores.

The formula for the median value would involve a number of symbols that would not be used again with exactly the same meanings.[1] In its place one may use the simple rule: *Count into the distribution to a point cutting off* $\frac{n}{2}$ *measures.* The point on the scale which best represents the division of the frequencies into two equal parts is the median or fifty percentile point. For the ten percentile or first decile point one would similarly give the rule: Count into the distribution to a point cutting off $\frac{n}{10}$ measures. For the first quartile or twenty-five percentile point, one would count into the distribution to a point cutting off $\frac{n}{4}$ measures. One would find other percentile points in a similar manner.

In adopting statistical methods from other fields, educational statisticians borrowed a formula for finding the median individual or score in a distribution, where such an individual or score actually exists. Of course, there is no median individual if there is an even number of individuals in the group. The formula directs one to *take as the median score the* $\frac{n+1}{2}$th *score.* If one had 25 scores, this formula would correctly indicate the 13th as the median score. If there were 26 scores in the group, however, this formula is useless, for it would direct one to take the 13.5th score, which is an absurdity, unless one understands it to mean the position between the 13th and 14th scores. Because of the great confusion which usually arises when students try to use this formula, and because the median

[1] T. L. Kelley, *Statistical Method*, p. 56.

score (when one exists) is the one containing the median value, the writer advises against learning and using the rule for finding the median score.[1] Learn to find the median point.

The formula for locating the exact mode is difficult to use, and the mode itself is so little used that no notice will be taken of it here. As was explained earlier in this chapter, the approximate mode or modal group is easily detected and needs no special rule.

Comparison of average measures. — The mode is the least useful of the three measures of central tendency or average. A relatively small number of extra individuals at some central place in the distribution may result in making that the mode, when under ordinary circumstances the mode would be in a different part of the distribution. The mode is difficult to determine accurately, and in many distributions there is no single well-defined mode.

The mean value is the most widely used measure of central tendency. It can be determined without arranging the data in any particular order, requiring only the addition of all the values and the division of their sum by their number. Even where the individual values of the original items are unknown, the mean may be calculated from their sum and their number. The size of each measure has a direct influence on the size of the mean value. This is undesirable in some instances, because a few extreme cases

[1] The errors which occur most frequently are the application of the $\frac{n+1}{2}$ formula even where there is an odd number of measures, and the use of it to find a point rather than a score. It is not correct to count into the distribution $\frac{n+1}{2}$ measures, as some educational workers have done.

may outweigh many normal cases in determining the size of the mean.

The median value is affected about equally by all the cases entering into its calculation. The extreme measures have no more weight than the normal measures. In fact, one does not need to know the sizes of the extreme measures in order to find the median value, if only he knows their number. The median is thought by many students to be somewhat more easily calculated than the mean.

Which measure of the central tendency of a distribution one should use in any particular case depends upon the nature of the measures and upon one's purposes. If the median and mean values differ greatly from each other on the same group of scores, both measures of central tendency should be reported, and perhaps the mode as well.

A shortened calculation of the mean. — Any one who expects to do a considerable amount of classroom testing should learn the following method of guessing at the mean value and correcting it later. The formulæ may be stated as follows:

$$M = G + d$$

$$d = \frac{\Sigma \text{ (algebraic) deviations from } G}{n}$$

In these formulæ, M = true mean value,
G = assumed or guessed mean value,
Σ = the sum of, and
n = the number of measures.

Stating this procedure as rules, one would have the following:

1. Guess the mean as nearly as possible.
2. Find the algebraic sum of the deviations or differences between the guessed mean and the original measures.

3. Divide the sum of these deviations by the number of measures.
4. Add the quotient to the assumed mean.

An example of how the guessed mean is to be corrected may be seen in Table XXXIX, in which the difference is found between each measure and the guessed mean (66.0). The plus and minus deviations are set down in separate columns to assist the eye in adding them. The algebraic sum of the deviations from 66.0 is + 1.8, and the number of measures is 6. The correction is then found by dividing

TABLE XXXIX

LAST SIX SCORES OF BOY IN HANDWRITING SPEED TEST

EXACT SPEED SCORES	G (GUESSED MEAN)	DEVIATIONS
64.9	66.0	−1.1
69.3	66.0	+3.3
66.8	66.0	+ .8
67.2	66.0	+1.2
64.1	66.0	−1.9
65.5	66.0	− .5
n = 6		Σ deviations = 5.3 − 3.5 = 1.8

1.8 by 6, as indicated in the formula. The quotient, .3, when added to the assumed mean, 66.0, gives a true mean of 66.3 letters per minute. In such a simple problem as this, of course, the guessing method is not actually shorter than the simple addition of the measures and division of their sum by their number would be. The more extensive the problem of mean determination, the more helpful it will be to be able to guess and then to correct.

TABLE XL

CALCULATION OF FIFTEEN-YEAR-OLD SPELLING MEAN

RANGE OF SCORES	DISTRIBUTION STEP	f (FREQUENCY)	DEV. FROM G (STEP 4)	f × DEV.	
22–25	6	14	+2	+ 28	
18–21	5	103	+1	+103	
14–17	4	161	0		
10–13	3	63	−1		−63
6– 9	2	26	−2		−52
2– 5	1	2	−3		− 6
		$n = 369$		Σ dev. = 131 − 121 = 10	

$d = 10 \div 369 = .027$ step; 4 words × .027 = .108 words;
Central value of step 4 = 16.0 words; $M = 16.0 + .1 = 16.1$ words

Table XL shows the calculation of the mean by guessing and correcting in the case of the spelling scores of Table XXXV. Step 4, or rather its central value, 16.0 words, is assumed as a guessed mean from which to note the deviations. The correction d is found to be .027 of a step, or .11 of a word. This correction, when added to the central value of the assumed mean step, gives the same mean value as was previously obtained by the longer method involving multiplications with larger numbers.

Summary. — Measures of the average or central tendency of a group may be reported exactly or they may be stated in terms of the score or distribution step in which they are located. The approximate mode or most frequent score is an easily determined measure of the central tendency of a distribution, but the exact mode is difficult to obtain

and is seldom used. The mean value of a group of measures is readily obtained and is probably useful in a larger number of ways than any other measure of central tendency. The median value is easily obtained and is not influenced by extreme cases to any greater degree than by the cases more nearly equal to it. Any one who wishes to state the mean score or step in a distribution of scores should first find the exact mean value and then report the score or step containing this value. The median score or step should likewise be located by finding first the median value and then the score or step containing this value.

This chapter supplies directions and formulæ for the determination of average values. Special attention is given to the mean, with directions for a shortened method of obtaining it by guessing an approximate value and correcting it later. The influence of the methods by which scores are assigned is emphasized again in this discussion of average measures.

QUESTIONS AND EXERCISES

1. A girl in the fifth grade was given a speed test in handwriting each day for two weeks. Scores assigned in the manner described on page 199 were as follows: 60, 63, 62, 64, 64, 65, 64, 67, 65, 64, 66, 63, 65, 66. Prepare a distribution of these scores and plot a surface of frequency to represent it.

2. Calculate the mean value and the median value of these scores. How do these values compare with those obtained by the fifth-grade boy described in the text?

3. What is the modal score? the median score? the mean score?

4. Calculate the mean value and the median value for each of the following distributions of scores for quality of handwriting:

Quality (Thorndike)	Frequency in Class			Quality (Ayres)	Frequency in	
	A	B	C		D	E
15	1	—	—	90	—	2
14	4	—	2	80	4	7
13	7	3	10	70	12	12
12	13	9	12	60	15	11
11	10	20	11	50	10	8
10	5	11	6	40	3	5
9	2	4	2	30	1	2
8	—	1	—			
Total	42	48	43	Total	45	47

5. Name the median score and the median step for each of the above distributions. Name the mean score and the mean step for each distribution. Name the approximate mode for each.

6. Calculate the mean value and the median value for each of the following distributions of scores in spelling:

No. of Words	Frequency in Class			
	F	G	H	I
25	—	1	—	—
24	—	—	—	1
23	—	1	—	4
22	1	—	—	7
21	1	—	1	10
20	4	3	2	9
19	7	2	3	6
18	5	—	2	7
17	3	1	—	4
16	3	3	—	3
15	2	1	4	—
14	—	1	2	—
13	—	1	1	—
12	—	—	1	—
Total	26	14	16	51

7. Calculate the mean value for the following distributions of scores in spelling a list of 80 words, employing the step given at the foot of each column as a guessed mean:

Distribution Step	No. of Words	Frequency in Class			
		J	K	L	M
13	75–79	—	—	1	—
12	70–74	—	2	1	—
11	65–69	4	6	4	—
10	60–64	10	11	7	—
9	55–59	23	15	9	—
8	50–54	48	27	6	3
7	45–49	55	19	5	8
6	40–44	40	13	2	12
5	35–39	31	8	—	10
4	30–34	16	4	1	6
3	25–29	7	1	—	4
2	20–24	1	—	1	2
1	15–19	—	—	—	2
	Total	235	106	37	47
Use as guessed mean		7	8	9	5

8. Calculate the median value for each of the above distributions.

9. Name the modal step in each distribution. What is the median score in each? the mean score?

REFERENCES FOR READING

D. C. Jones, *A First Course in Statistics*, Chapters IV and V.
W. I. King, *The Elements of Statistical Method*, Chapter XII.
W. A. McCall, *How to Measure in Education*, Chapter XV.
H. O. Rugg, *Statistical Methods Applied to Education*, Chapter V.
E. L. Thorndike, *Introduction to the Theory of Mental and Social Measurements*, pp. 44–46.

CHAPTER X

DISTRIBUTIONS OF FREQUENCY

DIRECTIONS for tabulating scores to show the frequency of each size of score were given in Chapter V, along with fairly complete instructions for the plotting of areas or "surfaces" to represent such tables of frequency. In the discussion and illustration of these operations, distributions were made of the ages of several groups of pupils, and the resulting frequency tables were represented graphically by surfaces of frequency. Chapters VII, VIII, and IX contained distributions of spelling and handwriting scores, with an occasional surface of frequency to assist the mind of the reader in comprehending the total situation more clearly than would have been possible without such graphic pictures. The purpose of this chapter is to examine the nature of the distributions of scores obtained with classroom tests and to present certain general characteristics, consideration of which may assist in the interpretation of practically all standard test results.

General symmetrical character of distributions. — If the reader will turn back to Figure 10 (page 103), he will observe very little similarity between the two distributions there represented, except that each distribution is represented by rectangular blocks. Figure 11 on the next page, however, represents two distributions that are much

more nearly alike. Each of the surfaces in Figure 11 is high in the middle and progressively lower as one's eye moves to the right or to the left from this tall central division, the mode. The two surfaces are also similar in that the steps by which the bounding lines approach the base line are smaller the farther they are to the right or to the left of the center. A third point in which these two surfaces resemble each other is that they extend to the right somewhat farther than to the left of their centers. Figure 13 (page 109) has exactly the same general characteristics as were just noted in Figure 11, although the height of the columns in Figure 11 is relatively greater than the height in Figure 13. This similarity between the distributions of age scores in different grades and in different cities would be found to hold true quite generally, except where a very small number of pupils were concerned, as in Figure 10.

Figure 24 (page 162) shows the surfaces of frequency for the distributions of spelling scores found in each of the elementary school grades, IV to VIII, at Baltimore, Maryland. It will be observed that each of these five distributions is remarkably symmetrical. The chief difference in the shapes of these surfaces of frequency is that the lower grades have larger and higher surfaces than are possessed by the higher grades. This difference in the areas of the surfaces is due to the smaller number of pupils tested in the higher grades. Each surface becomes lower as it departs from its mode, rather slowly at first, then more rapidly, and then more slowly again as it approaches the base line. The adjective "bell-shaped" is frequently applied to the general appearance of the bounding lines

of such a surface of frequency, and if the right angles in these lines were smoothed out they would certainly present much the same appearance as the outline of a bell.

Section C of Figure 29 (page 201) is a surface of frequency for the speed scores of a boy in handwriting. Its general appearance and shape are not far different from those of the age and spelling score distributions discussed above. As a matter of fact, one finds all sorts of scores falling into this general type of distribution, whether they are scores from repetitions of the same test on the same individual, or scores from many individuals of the same group taking the same test. The scores in any group are likely to cluster around the mode, and the number of scores higher than the mode is likely to be matched on the other side of the distribution by an approximately equal number of scores lower than the mode. There are likely to be fewer and fewer scores as one departs farther and farther from the mode in either direction.

The distribution found by testing a group of school pupils is of the same general type as one finds constantly in measuring natural objects. If one were to measure the lengths of the leaves on a tree in midsummer and should then make a tabulation of the lengths found, the same symmetrical type of distribution would appear. If one were to measure the weights of a thousand oyster shells, the distribution of the resulting weights would probably be quite similar to that for lengths of leaves or for the spelling scores of sixth-grade pupils. The "bell-shaped" surface of frequency is in fact so frequently found in the measurements of natural objects that it has been called the "normal surface of frequency." Figure 30 represents the

distribution of heights found among 6194 men in England.[1] This is almost a perfectly normal surface of frequency. Although most of the measurable facts of nature tend to assume an approximately "normal distribution," there are usually special influences that throw an undue proportion of the

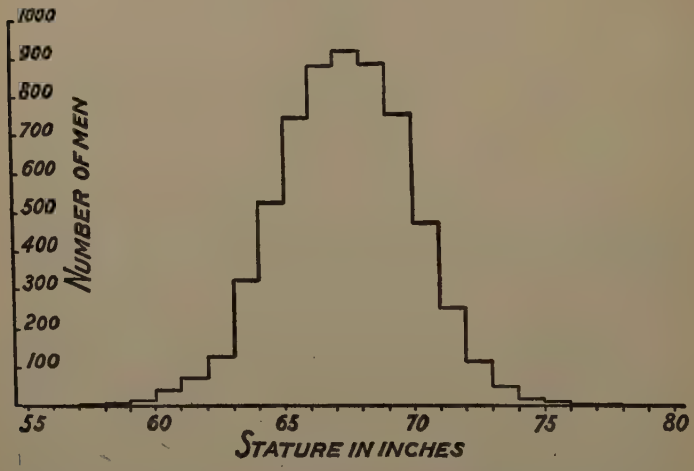


FIG. 30. — Surface of frequency showing heights of 6194 adult Englishmen. (After data quoted by Yule.)

measures in one direction or the other to such an extent that the resulting surface is not perfectly symmetrical.

Figure 31 shows two idealized normal surfaces of frequency, which differ from each other only in their compactness. Whether a surface of frequency is to be tall and thin, or short and wide, depends almost wholly on the selection of the unit for drawing it. If each score or measure is to be represented by a long, flat rectangle,

[1] G. U. Yule, *An Introduction to the Theory of Statistics*, p. 88.

the completed surface will tend to be low and wide. If each score is to be represented by a tall rectangle, the frequency surface will be relatively high.

The reader will probably realize without being told that smooth bounding lines, such as those possessed by the surfaces in Figure 31, are never actually obtained with tabu-

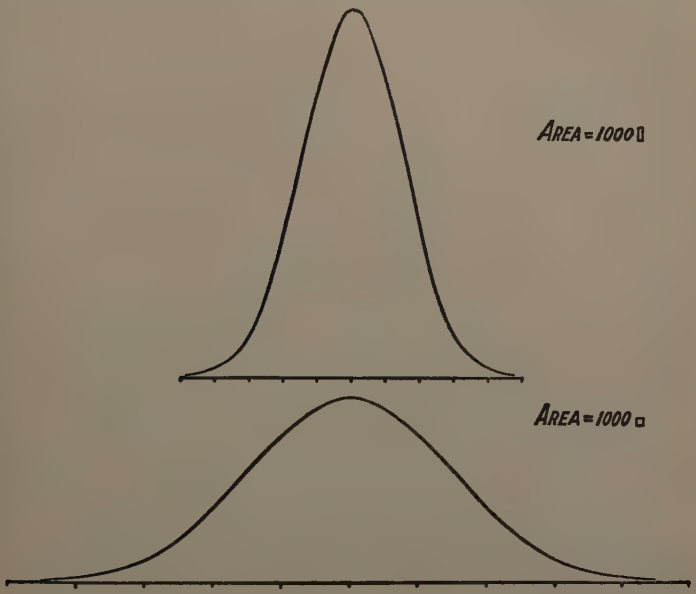


FIG. 31. — Normal surfaces of frequency. Both surfaces represent exactly the same facts and have the same total areas.

lated measurements. Perfect smoothness is most nearly approached when the number of measurements tabulated is very large and the divisions on the scale are very small. A bounding line with all the corners smoothed out as they are in Figure 31 is called a *frequency curve*. When such a smooth bounding line is drawn for a normal distribution

it may be called a *normal frequency curve*, or a *normal curve of probability*. The normal frequency curve and the normal surface of frequency are useful chiefly as standards by which to describe frequency curves and surfaces found by tabulating actual measurements. One may frequently shorten a description of a frequency surface by pointing out the respects in which it differs from a normal surface.

The effect of chance on distributions. — Whether a boy, who is taking daily tests of the speed of his handwriting, shall on a certain day make a score of 64, 63, 65, or 60 letters per minute, depends upon an enormous number of minor factors. His physical and mental condition may be disturbed at any second during the test by an infinite number and variety of noises, smells, visions, feelings, desires, memories, and other factors. One may fairly claim that, within certain limits, the exact score any pupil makes on any given test is almost a matter of chance. In order to understand the limits within which such chances usually work, it will be worth while to examine briefly the effects of chance on a distribution.

One of the most familiar examples of the action of chance is probably seen in the game of matching pennies. The chances that one player will match the other's coin are exactly equal to the chances that he will not match it, provided both are playing fairly. In the long run, therefore, neither player will gain a great advantage because his losses will almost exactly equal his gains. A coin which is fairly tossed will fall " heads " just about as often as it falls " tails."

If two coins are fairly tossed, chance will cause one-half of the throws to result in one head being up, one-fourth to

result in two heads, and the other fourth to result in no heads. The reason for this is clear when one examines the four possible combinations that may result from tossing two coins: *hh*, *tt*, *ht*, *th*. Both coins may fall heads, both may fall tails, the first may fall heads while the second falls tails, or the first may fall tails while the second falls heads. One of these combinations is as likely to occur as any other. There are two possible combinations having one head, but only one combination of two heads, and only one combination of no heads. In 100 throws of two coins, therefore, approximately 25 would yield no heads, 50 would yield one head, and 25 would yield two heads.

In tossing three coins, eight possible combinations might occur: *hhh; hht, hth, thh; htt, tht, tth; ttt*. Again one may assume that, if the coins are fairly tossed, one of the eight possible combinations is as likely to occur as any other. In 800 throws, therefore, approximately 100 would show no heads, 300 would show one head, 300 would show two heads, and 100 would show three heads.

When four coins are tossed fairly, any one of the following combinations is as likely to appear as any other: *hhhh; hhht, hhth, hthh, thhh; hhtt, htth, tthh, thht, thth, htht; httt, thtt, ttht, ttth; tttt*. In 1600 throws, therefore, the chances are that 100 will result in four heads, 400 will result in three heads, 600 in two heads, 400 in one head, and 100 in no heads.

The reader may, if he so desires, list the possible combinations of resulting heads and tails when larger numbers of coins are fairly tossed. He will find, for example, that with six coins there are 64 different possible combinations, distributed with reference to the number of heads as follows:

1 with no heads, 6 with one head, 15 with two heads, 20 with three heads, 15 with four heads, 6 with five heads, and 1 with six heads. As the number of coins is increased, it will be observed that the theoretically probable distributions maintain their perfectly symmetrical character, approaching more and more closely the "normal distribu-

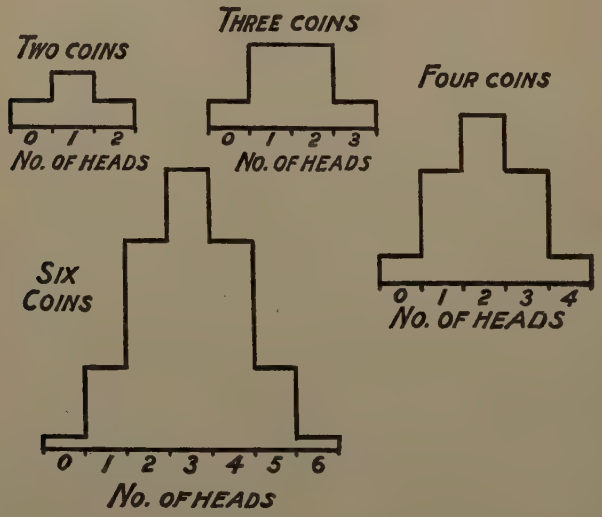


FIG. 32. — Theoretical results of tossing coins. Note the symmetrical distribution of the number of "heads."

tion." Figure 32 shows the most probable surfaces of frequency for tossing two, three, four, and six coins.

In tossing coins there are sometimes constant factors other than mere chance that affect the results in such a way as to disturb the perfect symmetry of the resulting distributions. Actual coin-tossing experiences of the writer are represented in Figure 33, as follows: A, one coin, 100 times; B, two coins, 100 times; C, four coins, 100 times;

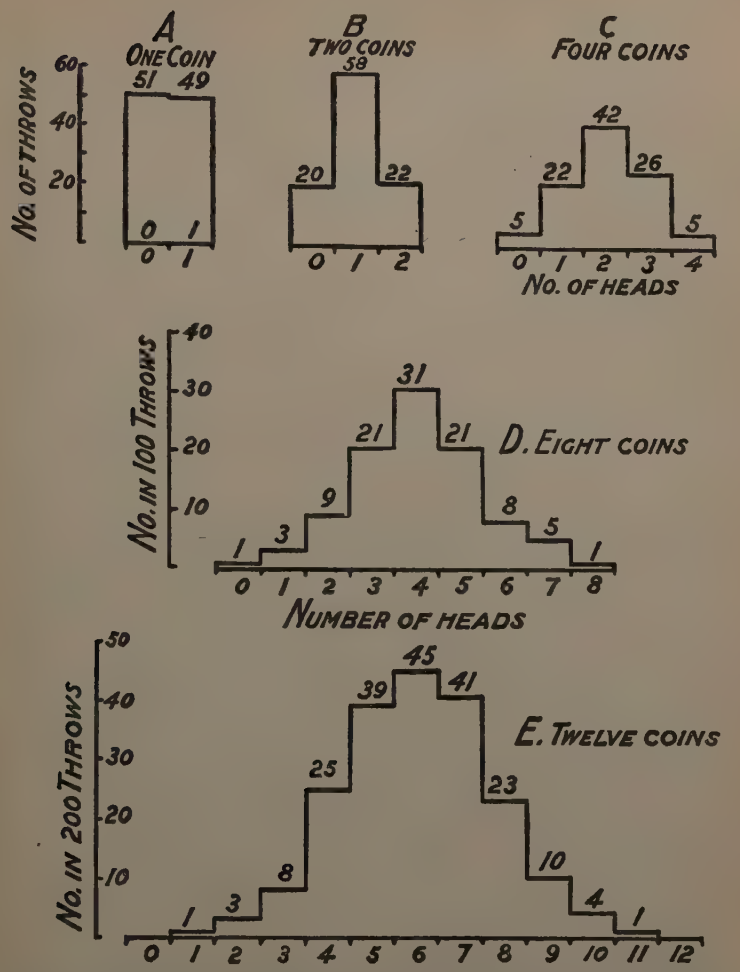


FIG. 33. — Actual results of tossing coins. These distributions are not as symmetrical as the theoretical distributions in Figure 32, but they approach the result of chance very closely.

D, eight coins, 100 times; and E, twelve coins, 200 times. In general, the actual distributions in Figure 33 are quite similar to the theoretically probable distributions in Figure 32. It is apparent from these figures that the favors of chance are distributed in an equitable manner and, when plotted graphically, approach the normal surface of frequency. An infinite number of coins tossed an infinite number of times would yield a surface of the form shown in Figure 31. Both nature and chance work according to certain definite laws which can be studied and depended upon.

The fact that the frequency surfaces obtained by tossing coins are very similar to the surfaces of frequency obtained by tabulating the scores of a group of pupils in a standard test, suggests the possibility that chance is about as powerful in determining the shape of the distribution in the case of test scores as it is in the case of coin tossing. By a careful analysis one may, of course, eliminate most of the accidental and unaccountable elements in what is usually called chance. Chance is merely a combination of many, small, independent, and sometimes unknown causes acting at the same time. In promoting pupils to a fifth grade, for example, the teacher has a general ideal of what fifth-grade ability should be, and most of the pupils she puts into the fifth grade are not far from her ideal fifth-grade pupil in their abilities. But her judgment is influenced by an enormous number of small items of dress, social status, physical health, moral character, and personal relations, as well as by many items of intellectual ability. A few of the pupils who are placed by the teacher in the fifth grade are not of fifth-grade ability. The teacher is likely to

overestimate ability just about as often as she underestimates it, and the number who are far on either side from the average or ideal is much smaller than the number who are near. The distribution of abilities in a fifth-grade class tends, therefore, to take the shape of the normal surface of frequency, for reasons similar in number, size, and independence to the reasons that cause coins to fall "heads."

One should not infer from the above discussion that every class of pupils, or even a majority of such classes, would be distributed "normally" if measured for any specific trait. The chief justification for expecting an approach to a normal distribution is the fact that the strength of mental traits in a given group of pupils is the result of a few dominant causes which affect all pupils in common, plus an almost infinite number of minor causes which are not common to all. The normal distribution of frequencies is given a large amount of attention in textbooks on statistical methods, chiefly in order that distributions which do not approach it closely may be accurately described by referring to the characteristics in which they differ from it. There is no wonderful power in nature which acts constantly to form normal distributions of measured facts. A perfectly normal distribution would be obtained in a set of educational measurements only as the result of chance.

Distributions of samples from a total population. — It is common to speak of the number of individual measures in a distribution or array as its "population," just as one speaks of the number of individual persons in a city or a state as its population. The size of the population tabu-

lated has a direct influence on the magnitude of the surface of frequency, but has relatively little influence on its shape and on its central tendency. The average value obtained from a distribution of 100 scores selected at random from the scores made by 1000 pupils would be almost the same as the average obtained from the scores of the total 1000 pupils.

It would not be necessary to rate the quality of each pupil's handwriting in order to determine the average quality for the 1000 fourth-grade handwriting samples that one might collect in a city school system. After shuffling the papers thoroughly, one could begin at the top of the pile and extract every tenth paper, which would furnish 100 samples to be rated and tabulated. The average quality calculated from such a tabulation would be practically the same as the average quality calculated from a tabulation of the total 1000 papers. The various quartile and decile points in the distribution for the 100 sample papers would not differ greatly from the corresponding quartile and decile points in the distribution for the total 1000 papers.

The principle of sampling is well illustrated by the miller or grain-buyer in purchasing wheat or other grains from the farmers. A farmer drives onto the scales at the mill or elevator with a load of wheat, the total weight of which is carefully obtained. The wagon is then unloaded slowly, while the miller catches a measure of the grain by dipping in here and there throughout the flowing stream, securing a random sample of the entire load. The sample is then carefully examined both for quality and weight. It is assumed that this measure of grain, obtained by a

random sampling of the entire load, is of the same average quality and of the same average weight as the remainder of the load. The farmer is then paid for the entire load on the basis indicated by the quality and weight of the grain in the sample.

The making of this type of random sampling, from which the main characteristics of a total distribution may be roughly calculated, is often a great economy of time and energy. If the sampling is perfectly random, representing each section of the total distribution in its proper proportions, most of the general facts might be learned from the distribution of a relatively small number of samples almost as well as from a distribution of the total number of ratings. If the sampling is not truly random or is inadequate in size, a false impression of the total situation might be conveyed by it. Such a false impression could only be corrected by taking a larger and more representative sampling, or by making a distribution of the total situation. Almost any reasonable number of samples, however, would give one a more reliable estimate of the average than could be obtained from a mere guess. Increasing the number of samples would, of course, increase the reliability of the estimate.

The number of random samples it is necessary to tabulate depends on the purpose for which the sampling is made, the range of the total distribution, and the accuracy with which one wishes to estimate the characteristics of the group. The effects of inadequate sampling may be observed in Figure 34, which shows the distribution for the number of heads turned up in tossing small numbers of coins a small number of times. In tossing two coins, for example,

Figure 34 shows that although five throws may give a very unsymmetrical distribution, ten throws are likely to yield a distribution that is very nearly normal. It also shows that even twenty throws were inadequate when tossing four coins, but that by the time the four coins had been tossed forty or fifty times the distribution began to have a symmetrical appearance. Thus, even where a total series of measurements forms an almost perfectly normal distribution, an inadequate sampling of these measures may yield an unsymmetrical distribution.

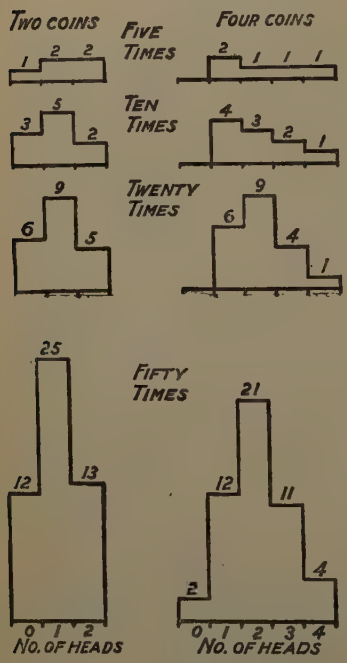


FIG. 34. — Effect of small number of cases on form of distribution. Note that tossing two coins ten times gives a fairly symmetrical distribution, while four coins must be tossed more than twice that number of times.

Skewness in distributions. — Not all distributions are symmetrical or normal, even when large numbers of individual cases are included. Figure 35, representing the handwriting scores of twelve-year-old pupils in the sixth-grade classes at Baltimore (Table XXVII), is an example of a distribution that is skewed positively, or toward the high end. A distribution is said to be skewed positively when its mean is higher in the scale than its mode.

A still less exact method of determining whether a non-

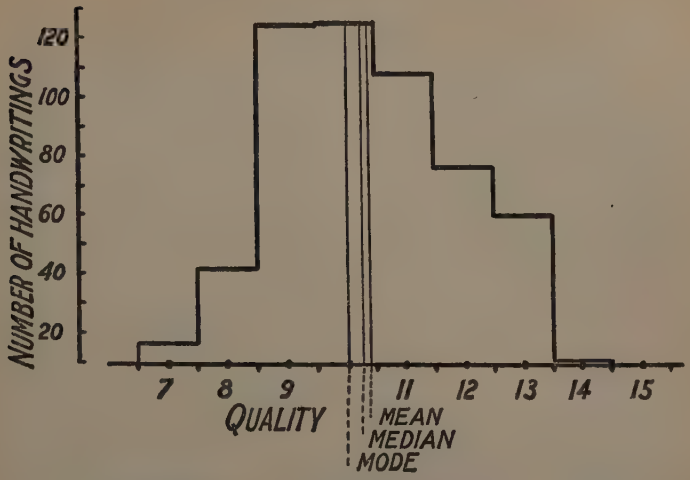


FIG. 35. — A distribution having positive skewness, — skewed to the high end. The mean is higher than the mode. The distribution is for the quality of twelve-year-old pupils' handwritings in Baltimore.

symmetrical distribution is skewed positively or negatively, is to look at its general shape and to determine by inspection whether the long slender side points up (to the right) or down the scale (to the left). In a distribution that is skewed positively, the long side of the figure extends toward the high end, while in a distribution skewed negatively, the long thin side points toward the low end of the scale. This method of determining by mere inspection the direction in which a distribution is skewed is too inaccurate for any but rough general use.

Figure 36 represents the spelling scores made by the seventh-grade pupils in Nassau County, New York.[1]

[1] M. R. Trabue, *Report of a Survey of Public Education in Nassau County, New York*, p. 212.

This distribution is skewed negatively, or toward the low end. The negative skewness in this distribution was probably due to the inadequacy of the spelling test used. Had there been a few additional and more difficult words in the list, it is probable that some of the pupils who had scores of 22, 23, 24, and 25 would have made higher scores. They were unable to make higher scores because there were no additional words to be spelled. It is sometimes possi-

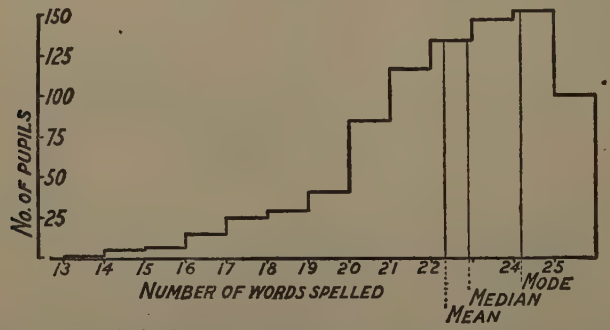


Fig. 36. — A negatively skewed distribution, — skewed toward the low end. The distribution represents the spelling scores of 853 seventh-grade pupils in Nassau County, New York. The test was of the same nature but was less difficult than the one described in Chapter VI. Many pupils would have made higher scores had there been more difficult words to be spelled.

ble to detect some such obvious cause for a skewed distribution, although in many cases the dominating influence which produces such a shape cannot be definitely determined. The method for determining the degree of skewness will be discussed in the next chapter.

Graphic representations of frequency surfaces. — The same facts are represented by Figure 37 as were represented in Figure 36. The only difference in the value of the two diagrams is that the individual scores in Figure 36 may be read much more readily, because of the greater definiteness

with which the different scores are set apart by the rectangular blocks. Although the areas of the two surfaces are equal, it is much easier to measure and compute the area from Figure 36 than from Figure 37. By connecting up the mid-points of each horizontal segment of the bounding line in Figure 36, one would have an exact duplicate of Figure 37. The name *histogram* has been applied to surfaces of frequency built up of blocks like Figure 36,

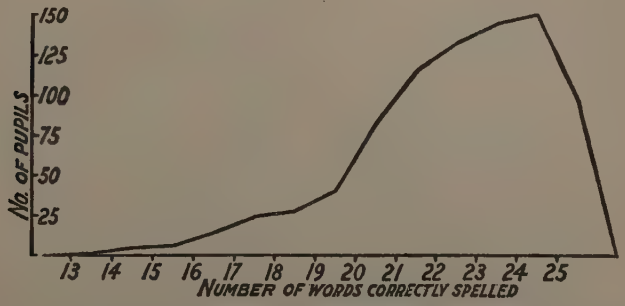


FIG. 37. — A frequency polygon. This type of frequency surface is more difficult to read than the histogram. Exactly the same spelling scores are represented here as in Figure 36.

while the name *frequency polygon* is ordinarily confined to surfaces drawn as in Figure 37. It is usually advisable to represent distributions by the histogram type of surface, because of the greater ease with which the facts it represents can be read.

Whether one is preparing a practice curve, a histogram, or a frequency polygon, it is economical to draw the figure with black India ink on white paper that has been ruled both ways with light blue lines. The light blue lines form squares or rectangles which assist one greatly in determining the height at which to place the bounding line or practice

curve above any given part of the base line. If the diagram is photographed and published, the blue lines will not show. In certain instances it is desirable for the reader to have the assistance of the small squares in reading the facts from the diagram, in which case the quadrille ruled lines should be of black or some other color than light blue. One should in all such cases see to it that the width and character of the lines he draws are quite distinct from the width and character of the ruled lines. It is usually most helpful to use quadrille or coördinate paper on which every tenth line is somewhat darker than the others, since this device helps one to count the squares quickly. If one's scale is by months and years, paper on which every twelfth line is heavier would be desirable, but for most of the work in educational measurements the decimal arrangement of lines will be found more convenient.

To illustrate the advantages of quadrille ruled paper, and to show how a rough irregular curve may be smoothed mathematically, Figure 38 has been drawn, representing in its heavy continuous line exactly the same facts as were given in Section *a* of Figure 29. There are in the continuous line so many short-time fluctuations that it is difficult to observe what degree of progress was actually occurring. The method by which these short-time variations are ironed out is essentially that of averaging the speed records of several consecutive days. The dashed line in Figure 38 shows the result of taking the arithmetic mean of the records for each group of three consecutive days.[1] The dotted line shows the result of taking the

[1] For daily scores see Table XXIX, p. 199.

Adding 59.4, the record for the first day, to 61.8 and 63.0, the

mean of the records for seven consecutive days, combining the records of each day with the records of the three days before and the three days after. As can be seen in this diagram, the larger the group of records averaged in the smoothing process, the more smoothly does the resulting curve run.

Exactly the same principles apply to the smoothing of the bounding line of a surface of frequency. Some pecu-

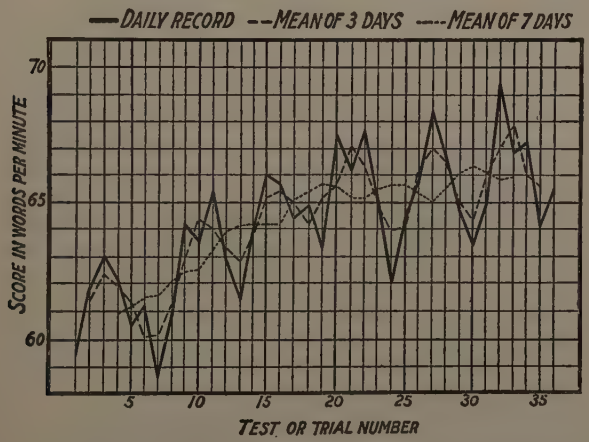


FIG. 38. — Smoothing irregular curves mathematically. The daily records of a boy's speed in handwriting averaged by groups of three and by groups of seven. For original scores see Table XXIX.

liarity in the method of scoring a test may result in every second or third step in the frequency surface being low. In order to smooth out this defect one could average fre-

records for the second and third days, respectively, and dividing by 3, gives 61.4 as the smoothed value for the second day. For the third day, 61.8, 63.0, and 62.1 have an arithmetic mean or smoothed value of 62.3 letters per minute.

quencies by groups of three steps, in a manner similar to that used in securing the dashed line of Figure 38. If the variations were over a larger range of the scale, one might iron them out by averaging frequencies by groups of five or seven steps, as in the case of the dotted line of Figure 38. As a general practice, it is better to deal with frequency distributions as they are found, without any effort to smooth them out, although the student may sometimes have occasion to eliminate minor irregularities in this manner. Whenever a surface has been smoothed, this fact should be noted in the text of the discussion as well as in the title of the drawing; one should not use it to misrepresent the data actually obtained in a standard test.

Summary. — Through the independent action of an immense number of small unrelated factors, measurements of a group of individuals in any given characteristic frequently approach a symmetrical distribution, which in its plotted form has come to be known as the normal surface of frequency. This type of surface is obtained by the action of chance, as in tossing coins or throwing dice. Other types of distribution may be described with reference to the ways in which they differ from the normal type. The type of the distribution and many of its characteristics can be estimated from a relatively small number of samples selected at random from the total group. As the number of cases included in the sampling becomes greater, the reliability of the sampling becomes greater. Absolutely accurate estimates can be made only by including all cases in the sampling studied. In plotting surfaces of frequency it is economical to use coördinate paper.

QUESTIONS AND EXERCISES

1. Mention one or two probable reasons for a lack of similarity between the two distributions in Figure 10.

2. In which direction are the distributions of Figure 11 skewed? Do you think of any facts which would have a tendency to skew them in this direction?

3. Repeat the experiment by which surface C in Figure 33 was obtained. Keep records for each group of ten throws separately. How do your results compare with those represented in Figure 34?

4. Outline an experiment (in tossing coins) by which you could determine approximately how many samples would have to be taken at random in order to secure a relatively normal distribution when the total range of the group was six steps. When the group ranged over ten steps.

5. Carry out one of the experiments outlined in 4. Compare the number of samples you find necessary with the number found necessary by some other experimenter. How would you explain the differences in your findings?

6. Redraw Section *c* of Figure 29, making a frequency polygon of it.

7. Draw a frequency polygon for Table XXVII (page 188). Superimpose a histogram representing the same distribution.

8. Smooth the distribution in Table XXVII by taking the mean of each three consecutive frequencies. Draw a histogram representing the smoothed distribution.

REFERENCES FOR READING

W. P. and E. M. Elderton, *Primer of Statistics*, Chapter III.

D. C. Jones, *First Course in Statistics*, pp. 55–62, 78–80.

E. L. Thorndike, *Introduction to the Theory of Mental and Social Measurements*, pp. 36–37, 44–46, 48–49.

G. M. Whipple, *Manual of Mental and Physical Tests*, Volume I, pp. 14–19, 30–31.

B. D. Wood, "The Reliability of Prediction of Proportions on the Basis of Random Sampling," *Journal of Educational Research*, Vol. IV, No. 5, pp. 390–395.

CHAPTER XI

MEASURES OF VARIABILITY AND DISPERSION

REPETITION of the same standard test with the same identical child may not give exactly the same result. If two persons were asked to measure the exact length of a room, it is not probable that their results would agree within a sixteenth of an inch. Even the length of the "standard yardstick" seems to vary slightly from one reading to the next, especially when the readings are made with microscopic precision. If two readings of the length of a bar of iron do not agree, how much less should one expect two readings of a boy's spelling ability to agree! The number and size of the factors that cause a child's mind to vary from one minute to the next are infinitely greater than those that cause a lifeless object like a bar of metal to vary. A living child should be expected to be more variable than an iron rod.

In order to know the variability of the measurements of an iron rod, or of a school pupil, one would need to secure an adequate sampling of its measurements. The factors which cause variability in an individual are so numerous and so independent of each other, that an adequate sampling of the measurements of any characteristic forms, when such measurements have been tabulated, a frequency distribution that is almost "normal." In trying to de-

termine the degree of variability for any measurement, therefore, one usually finds oneself dealing with a more or less normal distribution of samples of the measurement.

Not only does each pupil differ from himself from one moment to the next, but each pupil differs from every other pupil in the same group with him. Here again the factors which cause the differences are so many and so unrelated that an adequate sampling of them has a tendency, when

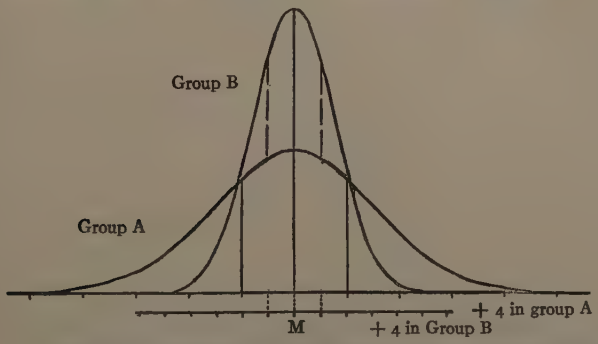


FIG. 39. — Differences in dispersion of two groups of pupils measured for general ability. Although both groups have the same average (M), the measure of dispersion (P.E.) in Group A is almost twice as large as in Group B. A class having a wide dispersion is difficult to teach.

tabulated, to form a more or less normal distribution. In one such tabulation there may be a relatively close grouping of the measurements, while in another the measurements may be widely spread. In other words, the dispersion may be small in one group and great in another. Whether one wishes to determine the degree to which measures for many individuals are similar, or to determine the variability of repeated measures for the same individual, the problem is essentially that of measuring the dispersion of a distribution of measurements.

Dispersion measured by defined ranges. — It is frequently quite convenient to be able to say that eighty per cent of the pupils made spelling scores between 10 and 13, or that half of the samples of a certain boy's handwriting fall between 8 and 10 on the Thorndike Scale. The range of the middle half of the measures is easily found by determining the 25 percentile value and the 75 percentile value; similarly, the range of the middle eighty per cent is found by determining the 10 percentile point and the 90 percentile point. The total range of a distribution is frequently reported but is not often helpful or significant, because it is practically determined by one or two unusually high or unusually low measurements.

The simplest measure of dispersion commonly used in accurate scientific work is the *semi-interquartile range*, which is exactly half of the range of the middle fifty per cent. Some writers have given this measure of dispersion the more easily spoken but less exact name *Quartile Deviation*. The abbreviation used to designate this measure is the capital letter *Q*. In determining the semi-interquartile range of a group of measures, one locates the lower quartile (25 percentile) and the upper quartile (75 percentile), and then takes half of their difference as the *Q* for which he is seeking. In Table XXX (page 200), for example, the 25 percentile is at 62.7 and the 75 percentile at 66.0, making 3.3 letters the range of the middle 50 per cent. The *Q*, or semi-interquartile range, is therefore about 1.7 letters $\left(\frac{3.3}{2} = 1.65\right)$.

Any number of other pupils might have exactly the same average score (64) as the boy whose scores are given in Table XXX, but some of them might be less consistent

than this one and others might be more consistent. The performance of this particular boy is quite well described by saying that his speed scores form a distribution that is only slightly skewed toward the low end, with a median value of 64.4 letters per minute and a Q of 1.7 letters. Another boy, the median value of whose scores was also 64.4, might have a Q of more than 2.0, indicating a larger variability, — that is, less consistency of performance. A third one might have the same median value but be more consistent, as indicated by a Q of perhaps 1.5 letters. The only fully complete record for any individual's performances would be a distribution of all his scores, but the next best record is a statement of the form of distribution with some measure of its central tendency and some measure of its variability. The Q is often entirely adequate as a measure of variability.

The distribution of scores for quality of handwriting given in Table XXXVIII may be described briefly as follows: an approximately normal distribution of 36 scores with a median of 50.8 and a Q of 8.47 (Ayres Scale). Interpreted in more popular language this description says that half of the 36 pupils wrote a better quality of handwriting than 50.8, on the Ayres Scale, while half of them wrote a poorer quality; and that the middle half of the pupils' writings were somewhere within 8.5 quality units of a common reference point.

Another group of pupils, having a median at 49.7 and a Q of 5.6 on the Ayres Scale, may be considered superior to the first group in a number of respects. The chief point in which the second group of pupils may be thought of as superior to the first is that their dispersion is less, — they

are more homogeneously grouped, more like each other in ability, and therefore more easily taught. Since the Q in the second group is only 5.6, the 25 percentile or lower quartile point must be at about 44.1 (49.7 − 5.6 = 44.1), which is slightly superior to the corresponding point (42.5) in the first distribution. And the probabilities are that the lowest score in the second is slightly higher than the lowest score in the more widely dispersed group. Such a situation is pictured in Figure 40, in which the heavy continuous line shows the frequency in the first group (Table XXXVIII), and the dotted line shows the frequency distribution in an imaginary group of the same size fitting the description: "Median = 49.7; Q = 5.6."

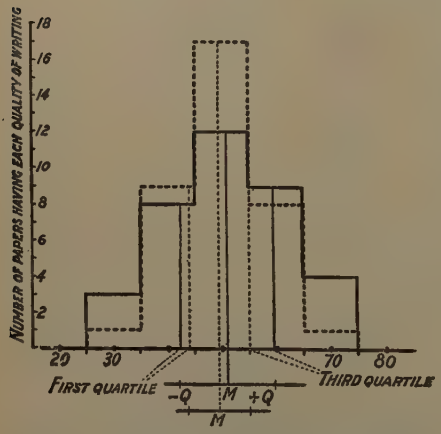


FIG. 40. — Distributions differing in dispersion and in average. Although the continuous line shows a distribution of handwriting scores (Ayres Scale) having a higher median (50.8), the total situation is less satisfactory, because the semi-interquartile range is greater (Q = 8.5) than in the distribution represented by the broken line (Median = 49.7, Q = 5.6).

The locations of the 25 percentiles do not always correspond exactly with the points 1 Q below the medians, and the point + 1.0 Q from the median is not always exactly the upper quartile point. This is because the Q is calculated by taking half of the difference between the upper

and lower quartile points, without regard to the location of the median point, and only when the distributions are symmetrical will + 1.0 Q from the median be the upper quartile point and − 1.0 Q be the lower quartile point. As a general practice the Q is used as a measure of dispersion only when the median point has been used as a measure of central tendency.

The median deviation. — In a perfectly normal distribution of measures, the *median deviation* of the measures from the median would be the same as the Q described above. Since few distributions are absolutely normal, one should learn to know the median deviation (Md. Dev.) and how to determine it. It is the median of the deviations of the measures from their central tendency, when all of these deviations are considered as positive quantities.

For example, eleven pupils were measured for quality of handwriting (Thorndike Scale), and the following array of scores was obtained: 9, 10, 10, 11, 11, 11, 11, 11, 12, 12, and 13. The median, mean, and mode of these scores are 11.0, which is the point from which deviations are to be calculated. Listing the deviation of each score from 11.0 gives: −2, −1, −1, 0, 0, 0, 0, 0, 1, 1, and 2. Considering all of these deviations as positive, one might rearrange them in order of size as follows: 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, and 2. Since there are eleven deviations, the median deviation will be the sixth in order of size; or, more exactly, the amount of deviation represented by the middle value of the deviation which follows the fifth $\left(\frac{n}{2} = 5\frac{1}{2}\right)$. The sixth in the array is a deviation of 1, which is the approximate median deviation of the eleven scores.

If one wishes to interpolate and thus secure a more exact value for the median of these eleven deviations, one must consider the possible range of a deviation of 1. With a central tendency at 11.0, a writing having a score of 12 would, as shown in Figure 41, have a true deviation from .5 to 1.5, depending on whether it were nearer x or y in the range allowed a score of 12. Similarly a score of 10 would deviate from 11.0 by from .5 to 1.5 steps, depending on the part of score 10 in which it was located. In the example considered above, there are four deviations of 1,

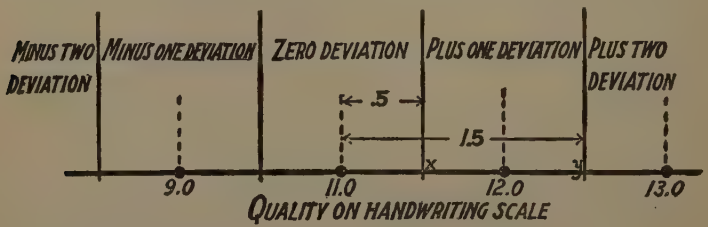


FIG. 41. — The meaning of a plus one deviation. A deviation of + 1 from an average score may be anywhere from .5 to 1.5 scale units, depending on whether the deviating score has a scale value of x or of y, respectively.

and one may assume that each covers an equal amount of the range from .5 to 1.5 of a step. To find the median value, one would need to choose the middle of the range allowed the lowest of these four. The middle of the lowest fourth of the range from .5 to 1.5 would be .625, which is the most exact possible calculation of the median deviation for the eleven scores considered.

The median deviation from the median is frequently known as the *probable error*. In a classic footnote,[1] Professor Thorndike remarks, "P. E. stands for Probable

[1] *Mental and Social Measurements*, p. 40.

Error, the traditional, but very misleading, name for the median deviation, which is not specially probable and not an error at all." The old name " probable error " may be traced to the fact that the median deviation was long used by biologists and other scientists in estimating the true average of certain measurements from limited samplings. In reporting the number of petals on a new flower, for example, a botanist might give the average number for all the specimens he had been able to collect, and then give the P. E. (median deviation) of his observations as an index of how variable the new flower seemed to be with respect to the number of petals. If any one happened to find a single blossom of this new variety, there would be approximately 50 chances in 100 that the number of petals would be within 1 P. E. of the average reported by the botanist who had first described it. Any variation from type or average might in this sense be called an error, although half of such errors would be within one median deviation of the type.

To give further familiarity with the process of determining the median deviation (Md. Dev. or P. E.), one may consider the variability of the boy's speed scores in handwriting contained in Table XXX (page 200). The mean value of these scores was found to be 64.3, and the median value was found to be 64.4 letters per minute. One might choose to tabulate the individual deviations from 64.4, and to treat the measures as if all those pupils having a given score had registered at exactly the mid-value of that score. In this case, one would have 7 deviations of .1 ; 5 deviations of .9; 5 deviations of 1.1 ; 3 deviations of 1.9; and 16 deviations of more than 1.9 letters. The median of the 36 deviations would be between the 18th and 19th

in order of value. Since both the 18th and 19th have a value of 1.9 letters, 1.9 may be taken as the approximate P. E. or Med. Dev. for the boy's speed tests.

One might, by making certain assumptions and interpolations, secure a more exact estimate (1.67) of the P. E., but it is hardly worth while to interpolate when there are so few cases in the distribution. As a matter of fact, it has become quite customary to secure the probable error indirectly by multiplying another measure of dispersion, the root-mean-square deviation, sigma, by the constant .6745, rather than to calculate the P. E. directly from the distribution. This method will be made clear in a later section of this chapter.

The mean deviation. — As its name suggests, the mean deviation is obtained by adding together all the deviations (regardless of algebraic signs) and dividing their sum by the number of measures. The mean deviation is the mean value of the deviations, in the same sense that the median deviation is the median value of the deviations. Because the extreme cases in a distribution are given relatively more weight in calculating the mean and the mean deviation than they are given in calculating the median and the median deviation, the mean deviation is always larger than the median deviation. In a perfectly normal distribution, the mean deviation is 1.184 the size of the median deviation; or, reversing the equation, the median deviation (P. E. or Md. Dev.) is .845 of the mean deviation (Mn. Dev.).

Taking as an example the eleven handwriting scores used to illustrate the calculation of the median deviation, one finds that $2 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 1 + 1 + 2$ equals 8; which divided by 11 equals .73, the mean devia-

tion of the eleven scores from their central tendency. This seems, at first glance, to be a much simpler calculation than that for the median deviation. It is also true that one may calculate the mean deviation without going through the labor of tabulating the scores. If one has the mean or other measure of central tendency from which the deviations are to be calculated, it is a fairly simple matter to note the amount by which each score deviates from the central tendency, add these deviations, and divide by their number.

In the case of tabulated data, the process is not greatly complicated. Again, with the material of Table XXX as the data, the different operations are given in Table XLI. In entering the deviations from the mean, it must

TABLE XLI

CALCULATION OF MEAN DEVIATION FROM TABULATED DATA

WRITING SPEED SCORES		(x) DEVIATION FROM 64.3 (MEAN)	PRODUCT OF FREQ. AND DEV. (fx)
Size	(f) Frequency		
69	1	+ 5.2	5.2
68	1	+ 4.2	4.2
67	3	+ 3.2	9.6
66	4	+ 2.2	8.8
65	5	+ 1.2	6.0
64	7	+ 0.2	1.4
63	5	− 0.8	− 4.0
62	3	− 1.8	− 5.4
61	3	− 2.8	− 8.4
60	2	− 3.8	− 7.6
59	1	− 4.8	− 4.8
58	1	− 5.8	− 5.8
Total	36	Sum (disregarding signs) 71.2	

be remembered that a score of 69 means from 69.0 to 70.0, that 68 means from 68.0 to 69.0; that the mean or central value of a score of 64 is 64.5; and that the central value of 63 is 63.5 letters. Multiplying each amount of deviation by its frequency, and adding the products (regardless of algebraic signs), give 71.2, which divided by 36 yields 1.98 as the mean deviation of the speed scores in writing.

The scheme described in a previous chapter whereby the mean is first guessed and later corrected does not interfere with the calculation of the mean deviation from the mean, if certain corrections are carefully made, and if the guessed mean (G) has a difference (d) of less than 1 from the true mean (M). If the difference (d) is found to be more than one unit, it would be well to guess again, choosing another guessed mean (G) which will make the difference (d) less than 1. The corrections to be applied in finding the true mean deviation under such conditions are contained in the following formula:

$$\text{Mn. Dev. from M} = \frac{\Sigma z + [l \cdot d] - (n - l)d}{n}$$

in which m's = the separate measures or scores,

n = no. of m's,

M = true mean,

G = guessed mean,

$d = \frac{\text{sum (algebraic) deviations from G}}{n}$,

Σz = sum (regardless of signs) deviations from G,

l = no. of m's less than M, and

$n - l$ = no. of m's greater than M.

It will be recalled that the formula for finding the mean was

$$\text{M} = \text{G} + d.$$

To illustrate the use of the formula for obtaining the mean deviation when the mean has been guessed, one may consider the handwriting scores (Thorndike Scale) obtained by fourteen-year-old pupils in the eighth grade at Baltimore, Maryland.[1] The tabulation of these scores and the operations involved in calculating the mean deviation are given in Table XLII. The extreme left-hand column gives, as usual, the sizes of the scores, and the next column gives their

TABLE XLII

HANDWRITING SCORES OF FOURTEEN-YEAR-OLD PUPILS IN THE EIGHTH GRADE

Handwriting Quality	(f) Frequency	(z) Dev. from G = 11.0	($f \cdot z$) Products
15	1	− 4	4
14	4	− 3	12
13	18	− 2	36
12	39	− 1	39
11	57	0	0
10	49	− 1	− 49
9	40	− 2	− 80
8	12	− 3	− 36
7	1	− 4	− 4
	$n = 221$		$91 - 169 = -78$

Calculation of Mn. Dev.

$$M = G + d = 11.0 + (-.353) = 10.647$$
$$\Sigma z = 91 + 169 = 260$$
$$l = 49 + 40 + 12 + 1 = 102$$
$$n - l = 57 + 39 + 18 + 4 + 1 = 119$$
$$\text{Mn. Dev.} = \frac{\Sigma z + (l.d) - (n - l)d}{n}$$
$$= \frac{260 + (102 \times -.35) - (119 \times -.35)}{221}$$
$$= \frac{260 - 35.7 + 41.65}{221}$$
$$= \frac{265.95}{221} = 1.203$$

$$d = -\frac{78}{221} = -.353$$

frequencies. The distribution seems to have a distinct mode at quality 11, and one may guess that the midpoint

[1] *Baltimore Survey*, Volume II, p. 343.

of this step (11.0) would be a satisfactory mean to assume. The third column (z) shows the amount by which the midpoint of each step deviates from the guessed mean, and the fourth column ($f.z$) gives the product of each deviation multiplied by its frequency. In order to facilitate the finding of the algebraic sum of these products, the positive products are placed in one position and the negative products in another.

The algebraic sum of the deviations from the assumed mean, given at the foot of the fourth column, is -78, while their sum when signs are disregarded is 260. If the assumed mean (11.0) were the true mean, one could divide 260 by the number of scores (221) and have the mean deviation without further work. The true mean, however, is .35 less than 11.0, so one must subtract .35 from each of the 57 deviations previously called 0, and the same amount from each of the 62 deviations greater than 0, and to each of the 102 negative deviations one must add .35; which is exactly what the formula does. When these corrections have been made, the sum of the deviations to be divided by 221 is approximately 266, instead of the 260 obtained from the guessed mean. The mean deviation for this distribution of handwriting scores is therefore 1.2 steps or units of handwriting quality.

The mean deviation is practically always calculated from the mean as a point of reference, although the mean deviation for the median is somewhat smaller than the mean deviation from the mean. It is frequently easier to calculate the mean deviation from the mean than from the median, and it has become customary to think of the mean deviation as a measure of dispersion to be used in connection

with the mean as a measure of central tendency. There is still another measure, however, which is more often associated with the mean than is any other measure of dispersion.

Sigma, the root-mean-square deviation. — This measure of the dispersion of a group of scores eliminates the plus and minus signs of the individual deviations by squaring each one. The sum of the squares of the deviations is divided by the number of measures, and the square root of the result is taken as the measure of dispersion. A descriptive but clumsy name for this measure is the *root-mean-square deviation*. Either the mean, median, or mode may be taken as the point from which to calculate the individual deviations in finding the root-mean-square deviation. When the mean is used as the point of reference, the term *standard deviation* may be applied to this measure of dispersion. The abbreviation for standard deviation is either the small Greek letter sigma (σ) or the English capital letters, S. D.

Handwriting Quality	(D) Deviation	D^2
13	+ 2	4
12	+ 1	1
12	+ 1	1
11	0	0
11	0	0
11	0	0
11	0	0
11	0	0
10	− 1	1
10	− 1	1
9	− 2	4
	Sum =	12

To return again for illustrative purposes to the eleven pupils measured for quality of handwriting (page 257), one may very easily determine the deviations, square each one, find the sum of these squares, divide by 11 (the number of measures), and extract the square root of the result. The sum of the squares of the deviations is 12; which when divided by 11 gives 1.09; the square root of which is 1.04, the standard deviation, or σ, for these eleven scores.

Table XLIII gives the calculation of the standard deviation for the 36 speed scores in writing tabulated in Table

TABLE XLIII

CALCULATION OF STANDARD DEVIATION FROM TABULATED DATA

WRITING SPEED SCORES		(x) DEVIATION FROM 64.3 (MN.)	(x^2) (DEV.)^2	($f \cdot x^2$) FREQ. × DEV.
Size	(f) Frequency			
69	1	+ 5.2	27.04	27.04
68	1	+ 4.2	17.64	17.64
67	3	+ 3.2	10.24	30.72
66	4	+ 2.2	4.84	19.36
65	5	+ 1.2	1.44	7.20
64	7	+ 0.2	.04	.28
63	5	− 0.8	.64	3.20
62	3	− 1.8	3.24	9.72
61	3	− 2.8	7.84	23.52
60	2	− 3.8	14.44	28.88
59	1	− 4.8	23.04	23.04
58	1	− 5.8	33.64	33.64
Total	36			$\Sigma f \cdot x^2 = 224.24$

XXX and used frequently to illustrate methods of calculation. This distribution is an example, not only of a situa-

tion where a score of 60 means from 60.0 up to 61.0, but also of a distribution of scores made by the same individual at different times. The column at the extreme right gives the products obtained by multiplying the square of each deviation by its frequency. At the base of this column appears the sum of the squares of all deviations, 224.24; which when divided by 36 gives 6.229, the square root of which is 2.49, the standard deviation for this distribution.

This makes the fourth measure of variability calculated for this distribution. These are given below, with a statement of the ratios between the four measures in a perfectly normal distribution. The standard deviation is the most important of the four and is always larger than the other three. The P. E. and the Q would, of course, be equal to each other in a normal distribution. It may be worth while for the student to remember that in a normal distribution the P. E. is about $\frac{2}{3}$ the size of the S. D., and that the Mn. Dev. is about $\frac{4}{5}$ the size of the S. D.

In distribution of speed scores:	*In a normal distribution:*
S. D. or σ = 2.49 letters	σ = 1.48 P. E. = 1.25 Mn. Dev.
Mn. Dev. = 1.98 "	Mn. Dev. = .798 σ
P. E. = 1.67 "	P. E. = .675 σ
Q = 1.65 "	Q = .675 σ

For distributions in which the mean has been found by guessing and correcting, the standard deviation may be readily found by the following formula:

$$\text{S. D. or } \sigma = \sqrt{\frac{\Sigma z^2}{n} - d^2}.$$

In this formula, as in that for finding the mean deviation, n is the number of scores, d is the algebraic sum of the individual deviations from the guessed mean divided by n,

and Σz^2 is the sum of the squares of the deviations from the guessed mean. The handwriting scores of Table XLII may as well serve to illustrate the method. It will be observed that the first four columns of Table XLIV are exactly the same as the four columns of Table XLII, and that the fifth column ($f \cdot z^2$) is merely the result of multiplying

TABLE XLIV

CALCULATION OF STANDARD DEVIATION FROM AN ASSUMED MEAN

HANDWRITING QUALITY	(f) FREQUENCY	(z) DEV. FROM G = 11.0	($f \cdot z$)	($f \cdot z^2$)	CALCULATION OF σ
15	1	+ 4	4	16	$\sigma = \sqrt{\frac{\Sigma z^2}{n} - d^2}$
14	4	+ 3	12	36	
13	18	+ 2	36	72	
12	39	+ 1	39	39	$= \sqrt{\frac{496}{221} - .1274}$
11	57	0	0	0	
10	49	− 1	− 49	49	$= \sqrt{2.2443 - .1274}$
9	40	− 2	− 80	160	
8	12	− 3	− 36	108	$= \sqrt{2.1169}$
7	1	− 4	− 4	16	$= 1.45$

$n = 221$ $\qquad$ $\Sigma z^2 = 496$ $\qquad$ $d^2 = \left(\frac{-78}{221}\right)^2 = .1274$

$91 - 169 = -78$

each number in the third column (z) by the corresponding number in the fourth column ($f \cdot z$). The sum (496) of all the squares of the deviations from the guessed mean, divided by the number of measures gives 2.24; from which must be subtracted the square of the difference between the

true and the guessed mean, which leaves 2.12 under the radical sign, or 1.45 when the radical is removed. The σ for this distribution is then 1.45 units. The mean deviation for this distribution was previously found to be 1.2 units of handwriting quality.

Uses of measures of dispersion. — One of the most common uses for measures of dispersion is to assist in describing a distribution of scores. The only complete description of any group of measures is a total distribution of them, but the next best, or in some cases even a more useful description would be a statement as to the form of the distribution, with some measure of its central tendency and some measure of its dispersion. For most of the standard educational tests, the median and the semi-interquartile range will tell the most important part of the story. For others it will be desirable to find the mean and either the standard or the mean deviation. If several groups have been measured on the same scale and have about the same central tendency or average, that one may be considered most homogeneous or unified which has the smallest Q or the smallest S. D. If entirely different scales are used in measuring two or more groups, however, their relative measures of dispersion will not in themselves give useful information as to which class is most closely grouped.

Another use that is frequently made of the dispersion measures is to indicate the importance of the differences between unequal scores made on tests. The size of a difference between two scores on a scale means little except in terms of the dispersion of the measures. Two inches is a relatively small difference in the heights of men, for example, but two inches difference in the lengths of their

thumbs would be an extremely large difference, because the variability of thumb lengths is so much less than the variability of total bodily lengths in the human race.

If the differences between the average scores of John and Mary in a reading test were 10 points and in a spelling test 7 points, one would not know until further information was available whether John and Mary were more alike in spelling or in reading. If it were known that a large class had taken both tests, however, and that the Q for the class was 5 in reading and 3.5 in spelling, one could say with some confidence that John and Mary were probably about as far apart in one subject as in the other (since in both subjects there was a difference of about 2 Q between them). If the σ of the class were known to be 5 in reading, and 5 in spelling, one could say rather confidently that John and Mary were more alike in spelling than in reading. On the other hand, if one knew that the class had a σ of 8 in reading and of 3 in spelling, it would be fairly safe to say that John and Mary were more nearly equal in reading than in spelling.

Teachers who try to assign scholarship marks from 0 to 100 do not all use the same range in the scale. One teacher's marks may have a standard deviation of 10 points on the scale where another in rating exactly the same students for exactly the same achievements may assign marks with a standard deviation of 5 points. The middle half of one teacher's marks may be between 70 and 80, while the middle half of another's marks may be between 80 and 90, and the middle half for a third teacher's marks may be between 65 and 85. In order to reduce all marks to a common basis, each may be converted into fractional parts of Q, or of σ,

above or below the average for the particular teacher's distribution of marks.

No one can tell, for example, whether a grade of 80% is good or bad, unless one knows the central tendency and dispersion of the marks assigned by the teacher to the entire group. One teacher, whose marks average 70% and have a σ of 10, would intend to compliment a student by assigning him 80%; while another whose marks average 85% and have a σ of 5 would intend to throw some doubt on the student's success by assigning him 80%. The one would be giving a $+ 1.0\ \sigma$ mark, while the other would be giving a $- 1.0\ \sigma$ mark, although 80% was the recorded mark in both cases. It is practically impossible to interpret any test score or any ability rating except in terms of the central tendency and dispersion of the total distribution of marks or scores.

Other important uses for measures of dispersion will appear in later chapters, particularly in the derivation and standardization of scale values for various test materials. It is highly worth while, however, for the reader to be able to calculate accurately at least the standard deviation, the mean deviation, and the semi-interquartile range. This chapter would also be incomplete without some discussion of how to measure the amount of skewness in a distribution which is not exactly " normal."

Amount of skewness in a distribution. — The " normal " distribution has been described, and the reader has been told that a distribution is skewed toward the high end when the mean is above the mode, and toward the low end when the mean is below the mode. There are all degrees of skewness, just as there are all degrees of attractiveness or of

spelling ability, and it is sometimes desirable to be able to describe a distribution in terms of the degree as well as the direction of its skewness.

It is undoubtedly clear, without explanation, that the amount of skewness must be in terms of the dispersion of the distribution rather than in terms of raw scores. Professor Karl Pearson, who has contributed many other useful methods of statistical treatment, has suggested that the ratio of the difference between the mean and the mode to the size of the standard deviation be used as a measure of skewness.

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}.$$

Because of the difficulty of calculating the mode with great exactness, and because the mode is usually about three times as far from the mean as is the median, Pearson's formula may for practical purposes be stated as follows:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\sigma}.$$

To illustrate the use of this formula, the reader may take again the speed scores in writing. The standard deviation, worked out in connection with Table LXIII, is 2.49 letters. The mean is 64.3, and the median is 64.4, which makes a difference of .1 letter per minute between these two measures of central tendency. Substituting in the formula $\text{skewness} = \frac{3 \times .1}{2.49} = \frac{.3}{2.49} = .012$, which is an almost negligible amount of skewness.

A distribution having a greater amount of skewness is represented in Figures 36 and 37. The mean of the seventh-

grade spelling scores in Nassau County was 22.28 words, while the median was 22.77 words. The difference between these two averages was therefore .49 of a word. The standard deviation of the distribution was 2.2 words. Dividing $3 \times .49$ by 2.2 gives .668 as an index of the amount of skewness in the distribution. Figures 36 and 37 may be described briefly as having a mean value of 22.3 words, a negative skewness of .67, and a standard deviation of 2.2 words.

Summary. — A given individual varies in his ability from one test to the next, and a group of individuals tested at the same time differ from each other by certain amounts. Whether one wishes to describe the dispersion of a group of individuals or the variability of one individual when measured a number of times, the problem is that of measuring the dispersion of a distribution of scores. The dispersion of a group of measures may be described in terms of the limits necessary to include a certain proportion of them, or in terms of certain standard measures of dispersion calculated from the distribution. Directions are given in this chapter for finding the semi-interquartile range, the median deviation or probable error, the mean deviation, and the standard deviation. Skewness in a distribution is measured by comparing the difference between two measures of average with the size of the standard deviation.

QUESTIONS AND EXERCISES

1. What are the limits of the middle 90 per cent of the handwriting scores in Table XXVII? The middle 80 per cent? The middle half?

2. What is the semi-interquartile range of these scores?

3. Calculate the semi-interquartile range of the spelling scores for twelve-year-old pupils given in Table XIX.

4. Find the median deviation or P. E. for the spelling scores of these twelve-year-old pupils.

5. Calculate the mean deviation for the same scores; the standard deviation.

6. What is the mean deviation of the handwriting scores of Table XXVII?

7. Find the standard deviation for the same handwriting scores.

8. Calculate the amount of skewness in the same distribution of handwriting scores.

9. Using data given in the second table on page 229, find Q (semi-interquartile range) for each of the groups of scores obtained in Schools F and I.

10. Find the mean deviation and sigma for each of the distributions of scores made by pupils of Schools F and I. Which of these two groups do you conclude is the more homogeneous with respect to spelling results?

11. Find P. E. for each of the two distributions. Convert these values into sigmas by formula ($\sigma = 1.48$ P. E.) and compare with the sigmas obtained by calculation in 10. Why are there differences?

REFERENCES FOR READING

W. P. Elderton, *Primer of Statistics*, pp. 47–54.
T. L. Kelley, *Statistical Method*, pp. 70–82.
W. I. King, *Elements of Statistical Method*, pp. 141–166.
H. O. Rugg, *Statistical Methods Applied to Education*, pp. 149–180.
E. L. Thorndike, *Mental and Social Measurements*, pp. 46–51, 59–60.

CHAPTER XII

TESTS OF SPEED IN ARITHMETIC

EFFICIENCY in calculation has long been a sphere in which friendly competition was possible, and it was one of the most obvious fields in which to attempt the measurement of educational results. In our pioneer days a "ciphering match" was possibly not as popular as a "spelling bee," although it was a real honor to be the winner in a mathematical contest, for relatively few persons were qualified to compete.

Pupils and teachers have long been inclined to feel greater confidence in the grades assigned for classroom work in arithmetic than they feel in the grades for other subjects. While two teachers are not likely in an actual trial to give exactly the same mark to the same examination paper in arithmetic, they expect their marks to agree fairly well. No recognition is given by most teachers to the fact that examination problems are not usually equivalent in difficulty nor equally fair to all pupils.

Contribution of S. A. Courtis to arithmetic tests. — In order to secure an accurate rating of the mathematical abilities of a group of pupils, the degree of difficulty of each problem in the test list must be known, the examples must be equally fair to all pupils, and an equal opportunity must be given for their solution. Some efforts had been made to prepare standard lists of problems before 1910, but the

chief credit for developing satisfactory tests along these lines belongs to Mr. Stuart A. Courtis of Detroit, Michigan. Most of the superintendents and supervisors who are now making constant use of educational measurements in their schools began by giving the Courtis Tests to measure the speed at which pupils solve problems in the fundamental operations of arithmetic. Year after year the scores made in Detroit and other cities were tabulated, studied, and published by Mr. Courtis, and each year his reports were more widely read.

While others were attempting to measure a complex group of abilities, which they usually grouped under the title "Ability in Arithmetic," Mr. Courtis wisely confined his measurements to certain definite phases of these abilities. Instead of having many varieties of addition problems, he arranged a test having a large number of addition problems of exactly the same type and difficulty. For subtraction, multiplication, and division, problems of a uniform type and difficulty were similarly selected and arranged. Each of the four fundamental operations of arithmetic is represented in Series B of the Courtis Standard Research Tests by a page containing a relatively long list of uniform problems, preceded by such instructions as the following:

"You will be given (eight) minutes to find the answers to as many of these (addition) examples as possible. Write the answers on this paper directly underneath the examples. You are not expected to be able to do them all. You will be marked for both speed and accuracy, but it is more important to have your answers right than to try a great many examples."

The first page contains twenty-four addition tests such as these:

927	297	136	486	384	176
379	925	340	765	477	783
756	473	988	524	881	697
837	983	386	140	266	200
924	315	353	812	679	366
110	661	904	466	241	851
854	794	547	355	796	535
965	177	192	834	850	323
344	124	439	567	733	229

The second page contains twenty-four subtraction problems of the following difficulty:

107795491	75088824	91500053
77197029	57406394	19901563

The following are four of the twenty-five problems in multiplication contained on page three:

8246	3597	5739	2648
29	73	85	46

The fourth page contains twenty-four division problems. Below are three of these problems:

$25\overline{)6775}$ $94\overline{)85352}$ $37\overline{)9990}$

The reader should observe that in each of the four tests illustrated above the various examples are of exactly the same general character and difficulty, and that all pupils have exactly the same amount of time for solving them. The differences between pupils revealed by these tests are in the number of problems solved during the limited time allowed for such work. Mr. Courtis has attempted in

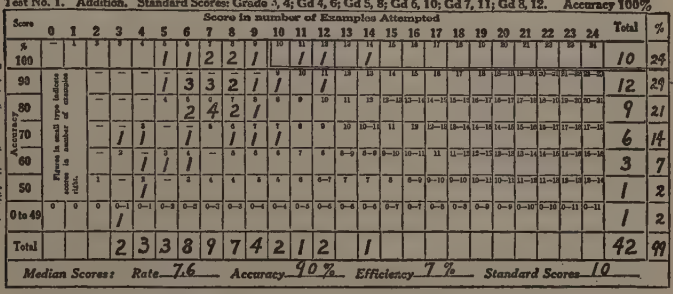
these tests to measure only two factors, the speed and the accuracy with which pupils can solve problems of a particular type and of a given degree of difficulty. The two scores recorded for each pupil are "No. Attempted ....." and "No. Right ....." In thus limiting his test to a definite field and controlling the factors surrounding that field, Mr. Courtis set a splendid example which has not always been carefully followed by other makers of tests.

Interpretation of Courtis test scores. — "Number attempted" means, if one has followed the printed directions, the number of problems for which the pupil has written complete answers. The teacher or examiner is directed by Mr. Courtis to have each pupil count how many problems he has finished before he gives up his paper.

"Count how many examples you have finished, and write the number in the score card in the corner under the number attempted. Do not count examples you have begun but have not finished. I am coming to your desk to see that you have written it in the right place."

A very ingenious Class Record Sheet, with a duplicate so perforated that it may be easily torn off and returned to Mr. Courtis, is supplied with the tests by the publisher. This record sheet, the addition part of which is reproduced with sample data, is a table so arranged that one may indicate the number of problems attempted by horizontal distance to the right, and the proportion of the problems correct by perpendicular distance from the base line. The small Arabic figures in the top of each rectangle show number of problems correct. One is to enter, in each rectangle, the frequency with which pupils (who attempted the number of problems indicated in boldface at the top of the column) solved correctly the number of

problems indicated in small type at the top of the rectangle. The boldface numbers at the left indicate the approximate percentage of accuracy of those pupils whose scores have been entered on that particular horizontal line.



Test No. 1. Addition. Standard Scores: Grade 3, 4; Gd 4, 6; Gd 5, 8; Gd 6, 10; Gd 7, 11; Gd 8, 12. Accuracy 100%

Score	Score in number of Examples Attempted																									Total	%
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24		
% 100	–	1	2	3	4	5	1	1	2	1		1	1		1											10	24
90				–	–	–	1	3	3	2	1	1		1												12	29
80				–	–	–	4	2	4	2	1	0	9	10	11	12	13–14	14–15	15–16	16–17	17–18	18–19	19–20	20–21	21–22	9	21
70				–	1	1	–	1	1	1	1	0	9	10	10–11	11	12	12–13	13–14	14–15	15–16	16–17	17–18	18–19	19–20	6	14
60				–	2	1	2	1	1	–	0	0	7	8	8–9	9–10	10–11	11	11–12	12–13	13–14	14–15	15–16	16–17	17–18	3	7
50			1	–	0	1	–	0	0	0	0	0	0–7	7	7	8	8–9	9–10	9–10	10–11	11–12	11–12	12–13	13–14	14–15	1	2
0 to 49	0	0	0	0–1	0–1	0–2	0–2	0–3	0–3	0–4	0–4	0–5	0–6	0–6	0–6	0–7	0–7	0–8	0–8	0–9	0–9	0–9	0–10	0–10	0–11	1	2
Total				2	3	3	8	9	7	4	2	1	2		1											42	99

Median Scores: Rate 7.6 Accuracy 90% Efficiency 7% Standard Scores 10

FIG. 42. — Class record sheet for Courtis Research Tests in Arithmetic, Series B. See text for explanations.

In the *Rate* blank at the bottom of the table, one is to record the median or mean of the attempts summarized on the bottom line. In the *Accuracy* blank one is to enter the median or mean of the summary column at the extreme right. In the *Standard Scores* blank one is to enter the standard for the school grade, which may be copied directly from the top of the form itself. In the *Efficiency* blank one is to enter the percentage of the total number of pupils who obtained 100% accuracy and at the same time attempted the standard or a larger number of the examples. The following quotation from Mr. Courtis will help to explain what he means by the word "Efficiency."

"For the Courtis Tests, Series B, the following standards are proposed. Those who consider them too high or too low are free to adopt others as they may see fit, but hereafter 'Efficiency' will be understood to mean efficiency in terms of the following:

TENTATIVE STANDARDS[1]

JUNE (OR MID-YEAR) STANDARD INDIVIDUAL SCORES IN SERIES B, ARITHMETIC TESTS

GRADE	ADDITION	SUBTRACTION	MULTIPLICATION	DIVISION
3	4	5	0	0
4	6	7	6	4
5	8	9	8	6
6	10	11	9	8
7	11	12	10	10
8	12	13	11	11

The author of these tests has prepared many excellent devices for helping both pupils and teachers to see just what the situation is with regard to each pupil's speed and accuracy. Graphic charts, cumulative record cards, and supervisory graphs are provided to stimulate the pupil's interest and to increase his efforts in bringing his scores up to "standard."

Dangers in the use of speed tests. — Mr. Courtis in a footnote says, "Rate will be used hereafter in place of Speed to indicate that the thing it is desired to measure is the natural rate at which children work, and not how much speed they are capable of developing." Unless pupils work in this test at their maximum rates, however, the results have little significance except for the extremely rapid and the unusually slow pupils. Those whose abilities are about average may not be adequately compared on the basis of their scores, unless all pupils are known to have put

[1] The actual scores of classes vary from these standards as much as four examples higher or lower.

forth their best efforts. When one is not working under pressure, his efforts are modified by a surprisingly large number of small distractions, and his scores may not be taken as satisfactory evidences of how his rate compares with the rates of his fellows who were subject to a different set of distractions. If pupils are allowed in a speed test to follow their natural rates, the results signify much less the relative abilities of the average pupils than would the results of the same test when pupils were giving it their best efforts and attention. Conclusive results can be obtained from a rate test only when the pupils are working at the highest speed of which they are capable at the time.

One of the gravest dangers of a test that measures primarily the rate or speed at which pupils work is that false inferences may so easily be drawn regarding the significance of the results. It is true, of course, that one may reason incorrectly from the results of any test, but the dangers of misinterpreting rate tests are especially great. To quote Mr. Courtis again:

"Scores in standard tests are records of performance, not measures of ability. Ability can only be inferred. A test measures only what is done with the particular material under the given conditions. A child's scores are determined by a wide range and a large number of factors. The most important of these are: Heredity, Maturity, Training, Physical, Physiological, Mental or Emotional condition. Tests reveal effects, but do not tell anything about causes. Do not draw conclusions hastily. Study the children. The tests merely point out the children in need of study."

The above cautions are particularly applicable to rate tests, although they are important considerations in the interpretation of all measurements. Psychologists and

educators have not yet been able to agree as to just what constitutes *ability* in any mental function. It seems somewhat more sensible to the writer, however, to think of ability in the fundamental operations of arithmetic as being demonstrated more clearly by increasing success in solving more difficult problems than by increasing speed in solving problems of the same difficulty. This belief is similar to the belief that greater strength in a pupil's arm would be demonstrated by his success in lifting, at arm's length, *a greater weight* than his fellows could lift, more adequately than it would be demonstrated by his success in lifting a moderately heavy weight *a larger number of times* than his fellows could in eight minutes.[1] If this belief is correct, then ability can be more satisfactorily measured by a series of performances graduated in difficulty than by a series of performances of equal difficulty.

A great deal of explanation and instruction is required in order to secure significant improvement in a test of the graduated difficulty type, while relatively little time and drill are necessary in order to secure improvement in a speed test. One superintendent reported, for example, that his school had in a month's time made two years' progress in arithmetic. As a matter of fact what he had done was to drill the classes on problems like those in the Courtis tests, thereby improving their average scores from a point two years below the standard to a point equal to the standard.

Rapid improvement is practically impossible on a test graduated from easy to hard, probably because improve-

[1] For further discussion on this point see M. R. Trabue, "Completion Tests for Public School Use," *Fifteenth Yearbook of the National Society for the Study of Education*, Part I, p. 53.

ment on such a test requires the development in the nervous system of entirely new nerve connections and routes, while improvement in the speed of doing equally difficult tasks involves chiefly the exercise of nerve connections that have already been formed. For these and other reasons it seems to the writer unwise for any one to infer very much about a pupil's general ability to add from his score on an addition test of the type used in the Courtis series.

One should be careful when giving rate tests to preserve uniform emphasis and time-limits. There is some evidence that shortening the time in a test well graduated from easy to hard does not lower the scores very much,[1] but any variation or change in the time limits in a speed test is certain to affect the resulting scores. Similarly, one examiner may read the directions for these tests, putting special emphasis on " You are not *expected* to be able to do them *all*," and " It is more important to have your answers *right* than to try a great many examples." He would probably secure a result quite different from that obtained by the examiner who read exactly the same words to his pupils, but who emphasized the initial sentence, " You will be given (four) *minutes* to find the answers to *as many* of these examples *as possible*." It is really difficult for any two teachers to read the directions for the Courtis Tests with exactly the same emphasis.

Another danger which all tests involve is particularly

[1] See, for example, Mr. Courtis' statement to the effect that a two-minute time limit on a Completion Test Language Scale, usually allowed seven minutes, did not greatly reduce the resulting class scores. Stuart A. Courtis, *Measurement of Classroom Products in the Gary Public Schools*, New York General Education Board, p. 337.

present in the case of speed or rate tests. Whenever a supervisor comes to a teacher's room to give a test, it is only natural that the teacher should accept the test as a sample of the type of work expected. Frequently the teacher undertakes, after such a test, to improve the abilities of her pupils in the type of work on which they have just been tested. The danger of suggesting too strongly that pupils should develop great speed in arithmetical calculations is not a mere theoretical possibility. Many cases have come to the writer's attention in which other phases of arithmetic have been neglected, and in some cases other subjects have been slighted, in order to achieve high scores in these tests. The fact that one can in a short time secure a more convincing improvement in *rate* of calculation than in *power* to calculate leads many ambitious teachers to over-emphasize rapid calculation.

Class scores on Courtis Arithmetic Tests, Series B. — The standards quoted on page 280 are for the number of problems attempted by individual pupils. The present writer has always felt that, for purposes of inter-school comparisons, the number of problems solved correctly was more important than the number of problems attempted. Mr. Courtis, on the other hand, has become more and more emphatic in advocating that records be kept only of the number of problems attempted and the percentage of accuracy. It is probable that for instructional purposes all three types of record should be kept, but where comparison with other schools and cities is desired, a single figure to represent the scores of a group of pupils is desirable. The writer believes that the number of problems solved correctly when pupils are doing their best is the only sound

basis for such comparisons. He has therefore collected comparative scores on that basis.

One objection to making comparisons between the average number of problems attempted and the accompanying accuracy in one place, and the average number of problems attempted and the accompanying accuracy in another place, is that the two schools may have very different ideals with regard to accuracy. In one school the purpose of the teachers may be to teach pupils to work slowly at first but always with absolute accuracy, while in the other school the attempt may be to develop speed first and then to improve the accuracy. No one yet knows just how many points of speed one should allow for a point in accuracy, or how many points of accuracy are equivalent to one point in speed. It is as yet impossible to agree which is better, 8 problems attempted with 50% accuracy or 4 problems attempted with 100% accuracy. For these reasons, the following scores do not mention "attempts" or "accuracy."[1]

Because the tests in different places were given at different times during the year, all the scores have been corrected in Table XLV to show what the probable score would have been in each place, if the tests had been given in April. The corrections were made on the assumption that each

[1] The writer is by no means opposed to the analysis of scores and the determination of percentages of accuracy. Each teacher and supervisor will do well to make such studies of each pupil's papers, but such analysis will be for the purpose of determining how to change the situation revealed. To compare one school with another, only number of "rights" should be used; but to know each child's individual needs his "attempts" should be counted and his accuracy calculated.

TABLE XLV

MEDIAN NUMBER OF PROBLEMS CORRECTLY SOLVED IN COURTIS ARITHMETIC TESTS, SERIES B. (APRIL)

IN GRADE	IV	V	VI	VII	VIII
ADDITION					
Average School System	*4.7*	*6.0*	*7.2*	*8.2*	*8.8*
Baltimore, Md.	4.2	5.5	6.2	7.0	8.3
Boston, Mass.	5.2	6.7	8.3	9.0	10.5
Detroit, Mich.	4.6	5.9	6.7	7.6	9.8
Kansas City, Mo.	3.7	4.3	5.5	6.2	
New Orleans, La.	3.4	4.5	5.6	6.5	7.9
Seattle, Wash.	3.5	5.2	6.4	7.4	7.8
SUBTRACTION					
Average School System	*5.9*	*7.5*	*8.8*	*10.0*	*11.2*
Baltimore	5.8	8.3	9.9	10.5	11.7
Boston	6.4	7.7	9.6	10.3	11.6
Detroit	4.9	7.2	8.6	9.5	11.5
Kansas City	4.2	5.6	7.1	8.5	
New Orleans	3.5	5.7	7.5	8.9	10.7
Seattle	5.0	7.9	9.8	11.0	12.1
MULTIPLICATION					
Average School System	*4.2*	*5.6*	*7.1*	*8.2*	*9.3*
Baltimore	4.2	5.7	6.8	7.3	8.1
Boston	4.0	5.6	7.2	8.2	9.2
Detroit	4.6	6.2	7.8	8.7	9.5
Kansas City	3.6	4.6	5.9	7.9	
New Orleans	3.2	4.7	6.1	7.7	9.5
Seattle	4.0	5.6	6.9	8.3	9.1
DIVISION					
Average School System	*2.6*	*4.7*	*7.1*	*8.6*	*9.7*
Baltimore	2.6	5.1	6.7	7.6	8.7
Boston	3.0	4.9	7.5	8.9	12.1
Detroit	1.8	6.3	6.6	9.3	11.0
Kansas City	2.7	4.0	6.3	7.4	
New Orleans	2.0	3.3	5.1	7.3	9.7
Seattle	1.2	4.5	6.8	9.2	11.2

school had a nine-months' school term, and that each month of school added to the score of a class one-ninth of the difference between the actual score of the class and the score of the next lower class. The student should bear in mind, when making calculations to compare with these median scores, that a score of "five problems right" means from 5.0 to 6.0 problems right, and similarly for each score.

The writer regrets that he has not been able to secure adequate standard scores by ages for these tests, but the lack of age-scores is not so important a matter with rate tests as it is with power or increasing-difficulty tests. The speed test scores are relatively less dependent on maturity and completeness of understanding, while the power tests are more directly dependent on these factors and show relatively small effects from recency and extent of drill. As far as the writer is aware, the only published distributions of Courtis Test scores by ages within the elementary grades are in the Baltimore School Survey, Volume II.

The Cleveland Survey Arithmetic Tests. — The next chapter will discuss a power test (an increasing-difficulty scale) in arithmetic. In that discussion it will be shown that a large amount of progress must have been made in the fundamentals of arithmetic before a pupil is able to solve a Courtis (Series B) problem correctly. There is considerable question in the minds of intelligent teachers as to whether pupils below the fifth-grade (eleven-year-old) level should be expected to take these speed tests. Many think it more important for pupils in grades below the fifth to spend their efforts in learning to handle new combinations rather than in trying to perform the relatively diffi-

cult operations rapidly. It is at least apparent that if rate tests are to be given in the lower grades they should be less difficult than those presented in the preceding pages. It was with this necessity in mind that Dr. Charles H. Judd and Mr. Courtis devised a new series of rate tests for use in the school survey at Cleveland, Ohio.[1]

The Cleveland Survey Arithmetic Tests are "spiral" in character. Instead of addition problems of only one difficulty-level there are four distinct levels, each of which contains a considerable number of equally difficult problems. Similarly, each of the other fundamental operations is tested on several levels. These tests measure the rate at which pupils can add, subtract, multiply, and divide, just as the Courtis Tests do, but they measure each process at several different degrees of difficulty.

Set A (Addition problems) has 65 simple problems like the following, and a time limit of 30 seconds.

1	6	9	0	4	1	7	9	3	2	1	3	6
2	6	5	1	2	3	7	6	0	4	5	8	9

Set E (also Addition) consists of 16 examples with a time limit of 30 seconds.

5	2	9	2	6	1	4
2	8	8	8	3	4	6
2	8	0	5	4	2	5
0	5	7	0	8	5	3
4	1	6	6	8	4	4

[1] C. H. Judd, *Measuring the Work of the Public Schools*, Russell Sage Foundation, pp. 94–123.

Set J (Addition) has 14 examples and a time limit of 2 minutes.

7	9	4	7	2	9	6
5	2	5	1	9	6	9
4	4	8	9	4	2	6
2	8	1	4	8	4	7
6	2	4	3	5	7	0
0	7	8	2	1	1	4
5	5	5	8	5	3	3
1	3	1	5	2	9	7
8	6	3	2	4	2	1
3	1	9	7	3	3	6
2	4	6	7	6	8	0
9	8	3	1	7	5	6
9	8	5	9	6	5	6

Set M (Addition) has 12 examples with a time limit of 3 minutes.

7493	8937	8625	2133
9016	6354	4091	1679
6487	2783	3844	5555
7591	4883	8697	6331
6166	1341	7314	6808

Sets B and F are subtraction problems; Sets C, G, and L are multiplication problems; Sets D, I, and N are division problems; and Sets H and O are problems involving fractions. As in the case of the Courtis tests, two scores are to be recorded, the number of problems attempted and the number of problems solved correctly. Here again it

is worth while for the teacher to make an analysis of each pupil's paper to determine whether he needs drill in speed or whether he needs instruction in greater accuracy. For purposes of comparison with other schools or cities, however, the score to be tabulated is the number of problems solved correctly. Table XLVI presents the averages of the results obtained in the upper half of each grade, as published in the school surveys at Philadelphia, Pennsylvania; St.

TABLE XLVI

CLEVELAND SURVEY ARITHMETIC TESTS

Average Scores by Grade

(Means of the median scores in upper section of each grade in Cleveland, Grand Rapids, St. Louis, and Philadelphia)

	STANDARD IN GRADE						
	3A	4A	5A	6A	7A	8A	
Test A	13.3	17.2	20.6	23.8	26.1	28.7	Addition
Test B	8.5	13.1	15.3	18.5	20.9	24.9	Subtraction
Test C	7.0	12.0	14.5	16.2	17.6	19.3	Multiplication
Test D	7.1	12.2	15.0	17.4	19.4	22.1	Division
Test E	3.9	4.7	5.5	6.4	6.9	7.6	Addition
Test F	2.2	4.7	6.5	7.5	8.7	10.3	Subtraction
Test G	2.2	3.9	5.0	5.5	6.0	6.7	Multiplication
Test H	1.4	1.8	4.6	5.2	6.5	7.9	Fractions
Test I	.8	1.2	1.9	3.0	3.9	4.6	Division
Test J	2.1	3.1	3.8	4.4	5.0	5.8	Addition
Test K	—	3.4	5.0	6.8	8.6	10.5	Division
Test L	—	2.1	2.8	3.4	4.0	4.6	Multiplication
Test M	1.5	2.6	3.3	4.1	4.6	5.2	Addition
Test N	—	0.8	1.1	1.5	1.8	2.3	Division
Test O	—	1.0	1.3	2.9	3.8	4.6	Fractions

Louis, Missouri; Grand Rapids, Michigan; and Cleveland, Ohio.

The Cleveland Survey Arithmetic Tests deserve greater popularity than they have thus far received. The Cleveland series measures the rates of work of pupils at many different levels instead of waiting until a relatively high level has been reached. The Cleveland series is somewhat more useful in the analysis of what each pupil's particular difficulty may be, mainly because it tests more elementary phases of ability than does the Courtis series.

It is probable that many schools have declined to employ the Cleveland series because of the relatively more complex sets of scores it necessitates and the relatively greater amount of labor it requires to tabulate the results. It is more satisfying to some people to have a brief result, which appears to indicate a general situation simply, than to have a complex result, which may contain much more complete information but be more difficult to explain satisfactorily. In spite, however, of its greater mental requirements of the examiner, the Cleveland series will supply more adequate information to the teacher and supervisor who are willing to study the results, than can be supplied by the original Courtis Tests.

One must not overlook the fact that the Cleveland tests are subject to great errors where the conditions and time limits are not carefully observed. Two or three seconds too much or too little will not disturb results so much in the Courtis Tests, where the total time is from four to eight minutes per test, but two or three seconds out of thirty seconds may cause quite a difference in the result. Greater care and training are therefore required in giving the Cleve-

land tests as well as in their interpretation. It would perhaps have been better had each of the Cleveland sets contained three or four times their present number of examples and allowed three or four times the present amount of time. In such tests the errors due to starting and stopping would have been less important. Perhaps such a set of tests will soon be developed and standardized.

Summary. — Teachers of arithmetic undertake to develop the abilities of their pupils in two directions, — to solve more and more complex problems, and to perform the fundamental operations with greater rapidity. In developing the first ability, careful explanations and demonstrations must be given, while in developing automatic control and speed a considerable amount of drill and practice is necessary. The Courtis Arithmetic Tests, Series B, measure the results of drill and practice in solving relatively complex examples, while the Cleveland Survey Arithmetic Tests measure the same types of ability at lower levels of difficulty. Because they provide for the possibility of a more complete checking of an individual pupil's difficulties, the Cleveland series is possibly a more useful measure; but because of the complex picture it presents, it has not become as popular as the Courtis series. Any rate test of this type is open to the criticism that it tends to over-emphasize a kind of instruction which may easily be stressed too much. To the extent that speed in the fundamental operations of arithmetic is desirable, however, these tests are useful measures.

QUESTIONS AND EXERCISES

1. How would you plan an experiment to determine whether it is easier to teach a rapid-but-inaccurate child or a slow-but-accurate child to be both rapid and accurate?

2. Would you be willing to have your efficiency as a teacher of arithmetic judged on the basis described on page 279? Why?

3. It has been found that most people, trying to do their best under experimental conditions, actually do more rapid and more accurate work when there are some distractions than when everything is quiet and comfortable. How would you explain this fact?

4. To what extent do you believe that speed in the fundamental processes of arithmetic should be emphasized in each elementary grade?

5. Determine the median scores in Addition made by the colored pupils of each age in Grades IV to VIII at Baltimore in the Courtis Tests, Series B, using the following distributions as a basis for your calculations.

Score No. Right	AGE								
	9	10	11	12	13	14	15	16	17
11	—	—	—	—	—	1	—	—	—
10	—	—	—	—	—	—	—	—	—
9	—	—	—	1	2	3	2	—	—
8	—	—	1	1	3	1	2	—	—
7	—	1	2	3	2	3	4	2	—
6	—	1	5	1	9	7	5	2	—
5	—	—	6	8	18	19	7	1	2
4	—	3	4	10	15	15	12	3	—
3	4	5	16	25	36	23	18	3	—
2	6	12	17	28	30	29	14	2	3
1	3	14	25	32	36	29	12	6	2
0	7	21	41	46	48	96	20	5	—
Total	20	57	117	155	199	226	96	24	7

REFERENCES FOR READING

E. J. Ashbaugh, "The Arithmetical Skill of Iowa School Children," Bul. No. 24, Univ. of Iowa, Extension Dept.

W. S. Monroe, "A Report of the Use of the Courtis Standard Research Tests in Arithmetic in Twenty-four Cities," Emporia, Kansas, State Normal School, Vol. IV, No. 8, New Series.

CHAPTER XIII

TESTS OF POWER IN ARITHMETIC

DOGMATIC statements should not be made regarding the exact nature of the ability measured by any given mental test, for the score obtained is quite certainly the result of a combination of several different abilities, each modified by numerous outside factors that prevent the score from being an accurate index of any one ability. One may, however, make a useful distinction between tests that measure primarily the strength and complexity of mental associations and tests that measure the rapidity with which certain of these associations operate. No one will deny that the ability to make mental connections quickly and with little effort is of great assistance in the making of a longer series of calculations. On the other hand, one may not successfully attempt to develop great speed in a given mental operation until after he has learned to follow the process without being prompted. The tests discussed in the previous chapter aimed to discover how rapidly pupils perform certain arithmetical operations at a given level of difficulty; while the tests to be discussed in this chapter were especially designed to measure how complex or how difficult an arithmetical problem pupils can solve correctly when given sufficient time.

The Woody Arithmetic Scales. — Like words in spelling, test questions in geography or arithmetic differ from each

other in difficulty. One example may be so easy that 80% of the pupils in a school can get the right answer in one minute, while another may be so difficult that less than 5% of the pupils can get the correct answer, even if an unlimited amount of time is allowed for its solution. Some problems are so advanced or so difficult that no third-grade pupil is able to solve them, although 75% of the seventh-grade pupils may be able to get the correct answers in less than two minutes. Every problem has its own peculiar elements of complexity, which when taken together give the problem as a whole a specific amount or degree of difficulty for pupils of a certain age or school grade.

Failure to recognize the degree of difficulty of different problems has frequently been responsible for false distinctions between students of equal ability in arithmetic, and for false estimates of the amounts of difference between the abilities of pupils who were not equally capable in this study. All of the problems in the Woody Arithmetic Scales have therefore been carefully examined and tested to discover just how difficult each one is for pupils of each grade in the public schools. It is possible with such information at hand to be fairly confident, for example, that John Smith, who solved only three of the ten problems on a given examination, is more capable in arithmetic than is James Black, who solved six of the ten problems correctly. This conclusion would be suggested if the three problems solved by John were each several times as difficult as the six solved by James. "How many problems were correctly solved?" is not so important in such cases as "Which problems were correctly solved?"

The Woody Scales are four in number, — one in addi-

tion, one in subtraction, one in multiplication, and one in division. However slowly a pupil may work, the time allowance on each of the scales is sufficient to allow him to solve practically all the problems of which he is capable. In Series B, which contains only a selected few of the problems to be found in Series A, the time allowed is 10 minutes for each page, while in Series A the pupils are allowed to work 20 minutes on each scale if they wish. Detailed instructions for giving the tests have been prepared and published by their author. Every teacher or supervisor who expects to use the scales in the classroom should secure a copy of these directions and should follow them exactly.[1]

The first six and the last two addition problems in the Series B scale are the following:

ADDITION

(1)	(2)	(3)	(5)	(7)	(10)
2	2	17	72	3 + 1 =	21
3	4	2	26		33
	3				35

(36)	(38)
2 yr. 5 mo.	25.091 + 100.4 + 25 + 98.28 + 19.3614 =
3 yr. 6 mo.	
4 yr. 9 mo.	
5 yr. 2 mo.	
6 yr. 7 mo.	

[1] C. Woody: *The Woody Arithmetic Scales — How to Use Them*, Teachers College Bureau of Publications.

The first four and the last three division problems in Series B are as follows:

DIVISION

(1)	(2)	(7)	(8)
$3\overline{)6}$	$9\overline{)27}$	$4 \div 2 =$	$9\overline{)0}$

(30)	(34)	(36)
$\frac{3}{4} \div 5 =$	$62.50 \div 1\frac{1}{4} =$	9)69 lbs. 9 oz.

The following quotation from Woody's directions may be given here:

"When all desired information has been secured, then say, without specially stressing any portion thereof, 'Every problem on the sheet which I have given you is an addition problem, an *and problem.* Work as many of these problems as you can and be sure that you get them right. Do all of your work on this sheet of paper, and don't ask anybody any questions. Begin.'"

The author of the tests supplies a list of the answers to be accepted as correct. He further directs that "In scoring the tests the standard for marking a problem correct is absolute accuracy, and, wherever possible, reduction to its lowest terms."

An elaborate method is provided for determining the scores of each grade or group on the scales of Series A, each score to be in terms of the degree of difficulty which would probably cause the class to register exactly as many successes as failures. This method is accurate and scientific, but in actual schoolroom practice few teachers are able to follow it successfully. Series B, samples of which are presented above, has therefore been used much more extensively than Series A. The score of each pupil is obtained

on one of the Series B scales, by counting the number of problems correctly solved. This simple method of determining scores is justified in this case, more than it would be in the case of the usual teacher's examination, by the fact that the examples are known to appear in the order of their difficulty, and by the further fact that the increases in difficulty between consecutive examples in Series B are fairly well equalized.

Table XLVII gives the difficulty value of each of the problems in Series B. These values were determined by the method described in Chapter XX of this volume, which assumes that the smaller the number of persons who are able to solve an example the greater its difficulty. The first problem in addition, for example, is given a lower difficulty value than the one numbered 7, not because it looks easier, but because the number of pupils in the second grade who can solve No. 1 is almost 50% greater than the number who can solve No. 7. It is probably true that any given group of pupils could be taught to recognize and to use the plus sign so well that No. 7 would for them become just as easy as No. 1, but under the conditions prevailing in the schools where these examples were standardized in the year 1916, No. 7 was much more difficult in the lower grades than was No. 1.

The reader should not interpret the above difficulty values as if they were calculated from an *absolute zero* point that indicated "just not any ability to add." Addition problem No. 5 has a difficulty of 2.83, but it is not for that reason twice as hard as No. 2 which has a difficulty of 1.40. No. 5 is, however, twice as far above Woody's *arbitrary zero* point in addition as is No. 2. Woody's zero point in

TABLE XLVII

DIFFICULTY VALUE OF EACH PROBLEM IN THE WOODY ARITHMETIC SCALE, SERIES B

ADDITION		SUBTRACTION		MULTIPLICATION		DIVISION	
No. of Problem	Value	No. of Problem	Value	No. of Problem	Value	No. of Problem	Value
1	1.23	1	1.06	1	.87	1	1.57
2	1.40	3	1.50	3	1.11	2	2.08
3	2.50	6	1.75	4	1.58	7	2.56
5	2.83	7	2.18	5	2.38	8	3.05
7	3.26	9	2.57	8	2.71	11	3.49
10	3.78	13	2.96	9	3.78	14	4.06
13	4.19	14	3.64	11	4.09	15	4.60
14	4.85	17	4.41	12	4.26	17	4.98
16	5.52	19	5.18	13	4.71	19	5.26
19	5.75	20	5.52	16	5.05	23	5.57
20	6.10	24	5.91	18	5.24	27	6.04
21	6.44	25	6.77	20	5.63	28	6.43
22	6.79	27	7.21	24	5.89	30	6.83
23	7.11	31	7.49	26	6.30	34	7.24
24	7.43	35	7.84	27	6.58	36	8.23
30	7.71			29	6.97		
33	8.18			33	7.29		
36	8.67			35	7.65		
38	9.19			37	8.02		
				38	8.53		

addition may possibly be several units above any absolute zero point in addition that would indicate " just not any ability." Furthermore, the zero point chosen for addition

is wholly different from and unrelated to the zero points in subtraction, in multiplication, and in division.

The difference between an absolute zero and an arbitrarily chosen zero point may be illustrated by a certain type of speedometer for automobiles, which does not register at all until the car is traveling at least ten miles per hour.[1] It would be foolish to say that the car had doubled its speed when such a speedometer moved up from an index of 1 to an index of 2, because the actual change in rate would be only from 11 (10 + 1 = 11) to 12 (10 + 2 = 12) miles per hour. It is perfectly legitimate in the case of such an arbitrary zero point, however, to say that an increase from 12 to 14 is equal to an increase of from 14 to 16, or is half as much as an increase from 12 to 16. Most of the scales devised for the measurement of educational achievements have an arbitrarily chosen zero point that is decidedly higher than any absolute zero, although in several cases the makers of the scales have tried to place it as near the true zero as possible. It is practically impossible to establish conclusively an absolute zero point in scales for mental processes.

Standard scores on Woody Scales, Series B. — It is unfortunate that as yet no adequate tabulations of the arithmetic scores made by the different age groups within each school are to be had. A given school may raise its grade scores by allowing only the unusually bright pupils to advance regularly from grade to grade, by eliminating from school the unusually dull pupils, or by placing the slow

[1] The writer is indebted to Professor G. D. Strayer for permission to use this illustration of the differences between absolute and arbitrary zero points.

pupils in special classes to which the tests are not applied, as well as by giving its pupils an unusually effective training in the fundamentals of arithmetic. On the other hand, a school which provides an unusually rich and varied course of study to meet the needs of all types of pupils will often retain larger numbers of its pupils and will therefore make relatively lower scores in each grade. The character and efficiency of the instruction may be higher in such a school than in the schools which eliminate their slow pupils. Tabulation of scores by ages within each school grade will make possible more accurate comparisons of the results of teaching, as distinct from the effects of retardation and elimination.

Table XLVIII shows the scores that have been obtained in a number of different school systems where Series B of the Woody Scales has been used. Although the first line, giving the average scores from a large number of places, is probably the best single set of scores with which to make comparisons, the seriousness of any amount of deviation from these standards can be interpreted in part at least by comparisons with the results from the specific places listed. It is worth while to note, for example, that the specific school systems differ relatively less from the average in these tests than individual schools differ from the average in the Courtis Tests in arithmetic. This is probably due to the fact that small differences in the amounts of drill make a much larger difference in rate scores than is made in power scores by differences in the amounts of instruction. The power scores are more fundamental and deep-seated than the speed scores.

The Stone Reasoning Tests in Arithmetic. — As was stated previously, in Chapter III, the Stone Reasoning

TABLE XLVIII

Comparative Scores by Grades, Woody Arithmetic Scales, Series B

Place	Date	Median Scores in Addition in Grades					
		III	IV	V	VI	VII	VIII
Average City	*Midyear*	*9.5*	*12.3*	*13.7*	*15.3*	*15.8*	*17.0*
Hoquiam, Wash.[1]	Mar., 1917	11.6	13.6	15.5	16.2	16.9	17.8
Seattle, Wash.[1]	June, 1918	—	11.5	15.0	16.2	17.2	17.4
Cherokee, Iowa	Feb., 1917	10.7	13.5	15.6	16.6	16.7	16.5
Webster City, Ia.	Dec., 1916	9.1	10.9	14.4	15.8	16.7	18.1
Denver, Colo.[1]	1918	9.8	13.2	13.0	14.5	15.6	15.7
Elyria, Ohio	Jan., 1917	10.0	12.8	13.6	14.5	16.0	16.0
Pittsburgh, Pa.	Jan., 1918	—	12.3	14.8	16.0	16.7	17.7

Place	Date	Median Scores in Subtraction in Grades					
		III	IV	V	VI	VII	VIII
Average City	*Midyear*	*7.3*	*9.3*	*10.7*	*12.6*	*13.7*	*14.2*
Hoquiam, Wash.[1]	Mar., 1917	9.3	10.2	11.8	13.8	14.7	15.0
Seattle, Wash.[1]	June, 1918	—	9.3	11.6	13.0	13.9	14.2
Cherokee, Iowa	Feb., 1917	7.2	9.6	11.2	13.6	13.8	13.8
Webster City, Ia.	Dec., 1916	6.8	9.0	11.1	12.7	13.9	14.4
Denver, Colo.[1]	1918	7.0	10.0	10.6	11.0	13.5	14.5
Elyria, Ohio	Jan., 1917	6.5	9.0	10.6	12.7	13.0	14.0
Pittsburgh, Pa.	Jan., 1918	—	9.4	12.0	13.2	14.1	14.9

Place	Date	Median Scores in Multiplication in Grades					
		III	IV	V	VI	VII	VIII
Average City	*Midyear*	*5.9*	*10.4*	*11.6*	*14.6*	*15.9*	*17.2*
Hoquiam, Wash.[1]	Mar., 1917	6.6	12.9	13.7	16.6	17.4	18.8
Seattle, Wash.[1]	June, 1918	—	10.4	13.5	15.6	17.0	17.8
Cherokee, Iowa	Feb., 1917	4.4	10.3	13.2	15.8	16.8	16.8
Webster City, Ia.	Dec., 1916	7.0	9.4	11.9	14.8	17.2	17.9
Denver, Colo.[2]	1918	6.5	10.4	10.8	13.8	16.5	15.8
Elyria, Ohio	Jan., 1917	6.0	9.0	10.0	14.7	16.0	17.5
Pittsburgh, Pa.	Jan., 1918	—	11.0	13.7	15.7	17.2	18.5

Place	Date	Median Scores in Division in Grades					
		III	IV	V	VI	VII	VIII
Average City	*Midyear*	*4.6*	*6.8*	*8.4*	*10.6*	*11.9*	*13.0*
Hoquiam, Wash.[1]	Mar., 1917	7.8	9.6	10.4	12.0	13.3	13.8
Seattle, Wash.[1]	June, 1918	—	—	9.8	11.1	12.4	13.0
Cherokee, Iowa	Feb., 1917	—	8.7	10.6	11.9	12.1	12.1
Webster City, Ia.	Dec., 1916	—	5.4	8.8	12.2	12.7	13.0
Denver, Colo.[1]	1918	4.5	6.4	6.5	9.2	11.5	12.0
Elyria, Ohio	Jan., 1917	5.0	7.7	8.0	10.0	11.2	12.0
Pittsburgh, Pa.	Jan., 1918	—	7.0	9.8	11.5	12.9	13.9

[1] Only the A (or high) section of the grade is here reported.

Tests in Arithmetic constitute one of the oldest standard tests for measuring the abilities of school pupils.[1] Stone's sixth-grade tests, first published in 1908, were in part the stimulus which caused Mr. Courtis to develop and improve his now widely-known tests. No textbook on elementary classroom tests should omit these pioneer tests. They still have a practical and scientific value, in spite of the progress that has been made during the past fifteen years in educational scale construction.

Dr. Stone emphasizes the following precautions to be taken when giving the Reasoning Test:

1. No announcement that a test is to be given should be made to the pupils.

"2. All directions to the pupils should be given by the tester. (If the tester is a person from outside the system and if the teacher and principal are busy filling blanks with helpful data, the original conditions will be more fully duplicated.)

"3. Principals and superintendents should refrain not only from communicating with pupils about the tests, but also from being present in the room either immediately before or during the test.

"4. The time limit should be kept exactly — fifteen minutes to the second.

"5. No mention should be made to the pupils of time limit.

"6. Furnish each pupil with a copy of the test and have printed side turned down, until all begin.

"7. Nothing should be said to pupils about the use of scratch paper, working in steps, amount of work to put down, etc."

The following are the words by which the examiner is to tell the pupils what to do:

"Take the materials that you usually take for your arithmetic work. Prepare two sheets of paper — headings and all. Have two sheets ready in case you may need them. *Use pencils.* Keep slip with printing turned down until we are ready to begin.

[1] C. W. Stone, *Standard Reasoning Tests in Arithmetic and How to Use Them*, Teachers College Bureau of Publications.

"Now, do you have everything ready? In order for you to do your best in this test, you will need to do just as all the other boys and girls who have taken the test have done. So pay close attention and do just as I ask you to do.

"You will not need to mark these (slips) papers at all. You will find directions at the top of these (slips) papers, telling you just what to do, so you will not need to ask any questions and — I do not think I need to say this to you, but I will, just because I have to all the other boys and girls, — be especially careful *not* to see anybody else's work. It is not easy not to see, but if you pay close attention to your own work only, the test will be the best. Begin."

Reasoning Test

Solve as many of the following problems as you have time for; work them in order as numbered:

1. If you buy 2 tablets at 7 cents each and a book for 65 cents, how much change should you receive from a two-dollar bill?

2. John sold 4 Saturday Evening Posts at 5 cents each. He kept $\frac{1}{2}$ of the money and with the other $\frac{1}{2}$ he bought Sunday papers at 2 cents each. How many did he buy?

3. If James had 4 times as much money as George, he would have $16. How much money has George?

4. How many pencils can you buy for 50 cents at the rate of 2 for 5 cents?

5. The uniforms for a baseball nine cost $2.50 each. The shoes cost $2 a pair. What was the total cost of uniforms and shoes for the nine?

6. In the schools of a certain city there are 2,200 pupils; $\frac{1}{2}$ are in the primary grades, $\frac{1}{4}$ in the grammar grades, $\frac{1}{8}$ in the High School, and the rest in the night school. How many pupils are there in the night school?

7. If $3\frac{1}{2}$ tons of coal cost $21, what will $5\frac{1}{2}$ tons cost?

8. A newsdealer bought some magazines for $1. He sold them for $1.20, gaining 5 cents on each magazine. How many magazines were there?

9. A girl spent $\frac{1}{8}$ of her money for car fare, and three times as much for clothes. Half of what she had left was 80 cents. How much money did she have at first?

10. Two girls receive $2.10 for making button-holes. One makes 42, the other 28. How shall they divide the money?

11. Mr. Brown paid $\frac{1}{3}$ of the cost of a building; Mr. Johnson paid $\frac{1}{2}$ the

cost. Mr. Johnson received $500 more annual rent than Mr. Brown. How much did each receive?

12. A freight train left Albany for New York at 6 o'clock. An express left on the same track at 8 o'clock. It went at the rate of 40 miles an hour. At what time of day will it overtake the freight train if the freight train stops after it has gone 56 miles?

Standard scores on Stone Reasoning Tests. — The author of these tests seems to be more interested in detecting the points of weakness in individual pupils than in making comparisons between schools or between classes. His method for marking or scoring the tests is to give the pupils credit for each problem in which the reasoning is correct, even though the calculations may be inaccurately done. Fractional parts of a credit are to be allowed in case only a part of the reasoning is correct, the size of the fractional credit depending on the number of reasoning steps correctly taken. From actual trials it was learned that not all of the problems are of the same difficulty. Each problem was therefore assigned a weight which would represent approximately its true value.

Those teachers and supervisors who undertook to give these tests and to score them as Dr. Stone directs, with fractional values for right reasoning and with irregular weights, had considerable difficulty. In the Butte (Montana) School Survey,[1] these tests were used for measuring the pupils in Grades V to VIII, but were scored according to the correctness of the answers rather than according to the correctness of a pupil's choice of method. This method

[1] G. D. Strayer, et al., *Some Problems in City School Administration*, pp. 170–184. It should be noted that Dr. Stone originally intended this test to be used only in Grade VI.

of scoring, and a simple method of weighting the problems have since been used so extensively that they may be reported here as standard.

In making comparisons with the scores in Table XLIX, therefore, each problem should be marked " right " if the answer is correct in every detail, or " wrong " if incorrect in any respect. Each of the last five problems (that is, problems eight to twelve inclusive) are to be given double weight in determining a pupil's score. The maximum score any pupil may obtain under this scheme is 17, while under the original method of scoring a pupil might obtain 17.2 credits if all his problems were reasoned correctly. Those who wish to study their own pupils' errors, rather than find

Problem No.	1	2	3	4	5	6	7	8	9	10	11	12	Total
Weight (to use)	1	1	1	1	1	1	1	2	2	2	2	2	17
Weight (Stone)	1	1	1	1	1	1.4	1.2	1.6	2	2	2	2	17.2

the comparative standings of their classes with classes elsewhere, should obtain Dr. Stone's booklet of directions and follow his suggestions carefully. Table XLIX gives the median results that have been obtained in various places by the simpler scoring method.

The approximate averages given in the last line of Table XLIX have been estimated by the present writer rather than by Dr. Stone. No supervisor or teacher need accept or use these estimates if he has other standards with which he wishes to make comparison. These have been calculated from reports in the writer's hands and are intended to serve merely as guides for those whose data are less complete.

Those who follow Dr. Stone's method of scoring may compare directly with the results reported in his discussion.

TABLE XLIX

MEDIAN SCORES BY GRADES IN STONE REASONING TESTS (BASED ON NO. RIGHT)

NAME OF PLACE	REPORT GIVEN IN	SCORE IN GRADE			
		V	VI	VII	VIII
Butte, Mont.	Survey, 1914	2.7	4.4	6.3	8.2
Columbia, S. C.	Survey, 1918	3.0	5.0	6.3	5.4
Elyria, Ohio	Report, 1917	2.9	4.9	6.0	8.0
Janesville, Wis.	Survey, 1918	2.4	3.4	5.5	6.3
Lead, S. D.	Report, 1915	3.2	5.3	7.6	9.6
La Porte, Ind.	Report, 1915	3.4	4.6	8.1	8.4
Nassau Co., N. Y.	Survey, 1916	—	4.5	—	7.2
Paterson, N. J.	Survey, 1918	—	5.6	—	9.0
Salt Lake City, Utah	Survey, 1915	4.3	6.9	9.1	11.0
San Francisco, Calif.	Survey, 1917	2.8	4.5	5.4	6.8
Approximate (middle-of-grade) Average		3.0	4.8	6.3	8.0

The Buckingham Scale for Problems in Arithmetic. — The Stone Tests were originally selected and standardized for testing sixth-grade pupils only. They constituted the first serious effort at arranging a set of problems in the order of their difficulty for testing purposes. It is not strange, therefore, that they are not as complete and satisfactory as could be desired. Dr. B. R. Buckingham has recently prepared and evaluated a large number of arithmetic problems, which are now available, in two forms or

series, for measuring the problem-solving abilities of pupils in Grades III to VIII.

The exact degree of difficulty of each example in the Buckingham Scale has been determined by extensive experiments and is printed on the test blank with the statement of the problem itself. Each series or form of the scale is broken up into three divisions, the first of which is for testing pupils in Grades III and IV, the second for pupils in Grades V and VI, and the third for pupils in Grades VII and VIII. Each division overlaps the one before it, however, so that all pupils are measured on exactly the same scale. The difficulty values in each division of Form I are as follows:

1st Division: 27–30–33–36–39–42–44–48–51–53

2d Division: 42–44–48–51–53–55–57–59–61–63–65–67–69–71–73

3d Division: 65–67–69–71–73–75–77–79–81–83–85–87–89–91–94

The following are selected problems from Form I of this Scale.

Value		
(30)	23 children belong to our class, but only 19 are present. How many children are absent?	ANSWER

(51) If an electric car runs 9 miles an hour, how many hours will it take to travel from one city to another, 117 miles away? ANSWER

(71) If 0.78 of the weight of potatoes is water, how many pounds of water are there in a bushel of potatoes, if a bushel of potatoes weighs 60 pounds? ANSWER

(91) On a map, a line $2\frac{3}{4}$ inches long represents a distance of 132 miles. How many miles are represented by a line $8\frac{1}{2}$ inches long? ANSWER

The publishers[1] supply with the tests a class record sheet on which a record may be made of each child's success

[1] The Public School Publishing Company, Bloomington, Illinois.

or failure on each problem. This enables the teacher or supervisor to add up and to enter at the bottom of the sheet the number of pupils in the class who were able to solve each particular problem. This summary of the scores in each class on each problem is one of the most helpful devices the educational measurements workers have ever suggested to supervisors. In the case of the Buckingham Scale it is unusually helpful. Since the general difficulty of the problems is known to increase gradually, any sudden drop in the number of pupils able to solve the next most difficult problem, will indicate that there is something in the character of the problem itself on which these particular pupils have not been well instructed. By glancing up through the individual records the teacher may quickly identify the particular pupils who need further instruction. This scheme of keeping the record of each problem as well as the record of each child is available, of course, to any teacher or supervisor with any examination. It has extraordinary value, however, with questions or problems which have been carefully graduated in difficulty, such as the Buckingham problems.

The score of each pupil is determined by the difficulty value of the hardest problem he has answered correctly. If the pupil has answered correctly all the problems up to a given point and none thereafter, his score is the value of the last problem rightly answered. If the pupil has failed on one or more examples and then solves a more difficult one correctly, his score is the value of the last problem correctly answered minus a certain amount for each previous problem wrongly answered. If the wrongly answered problems are in Division I, 3 points are to be subtracted for

each; while if they are in Division II or Division III, 2 points are to be subtracted for each wrong answer. For example, suppose a seventh-grade boy solved everything correctly up to and including No. 9 in Division III, but failed on all examples beyond that. No. 9, the most difficult problem answered correctly, has a value of 81, which is therefore the boy's score. If another boy solved No. 9, but failed on No. 1 and No. 8, and on all beyond No. 9, his score would be 77 ($81 - (2 + 2) = 77$).

It would be unusually valuable to have tabulations of scores on this test by age groups, but these are not yet available. The class record sheets supplied by the publishers give "Mid-Year Standards" dated November, 1920, and based on examinations of more than four thousand pupils in each school grade. These are given in Table L.

TABLE L

MID-YEAR STANDARDS IN BUCKINGHAM SCALE FOR PROBLEMS IN ARITHMETIC

IN SCHOOL GRADE	III	IV	V	VI	VII	VIII
75 percentile	43	52	65	72	85	89
50 percentile (median) .	39	45	61	66	80	84
25 percentile	34	42	55	61	76	79

Summary. — This chapter presents three standard arithmetic scales that measure how difficult an example can be solved correctly. The Woody Scales measure this power in the fundamental operations of addition, subtraction, multiplication, and division. The Stone Test and the

Buckingham Scale measure this power in the solution of written problems, in which the pupil must not only perform the operations accurately but must first select the operation to be performed. The Stone Reasoning Test has the distinction of being one of the oldest members of the standard scale family. The Buckingham Scale of Problems in Arithmetic is easily applied, quickly scored, and definite in its information. With all these measures of power to solve more and more difficult problems, it will be a decided advantage to tabulate scores by ages within each grade. Such tabulations will allow summaries by age as well as by grade and will correct false impressions regarding the quality of instruction now suggested by grade scores.

QUESTIONS AND EXERCISES

1. Would you classify the spelling test lists on page 136 as primarily tests of spelling speed or tests of spelling power? Which type of test is more important in spelling? Why?

2. What is the purpose of speed or rapidity in arithmetic calculations?

3. Would training in the method of solving more difficult problems have a great effect on the rate at which one could solve relatively simple examples? Would drill in the speed of performing simple operations have any appreciable effect on the power to solve more difficult problems? Which ability seems more complex? more fundamental? Why?

4. The method of scoring each of the tests described in this chapter is to "count the number of problems correctly solved." How nearly right (expressed in a percentage) could a solution be and yet be scored as "wrong" or "incomplete"? What then are the possible limits of a score of 10?

5. To what extent do the Woody Scales measure rate or speed?

6. Is zero on the Fahrenheit Scale absolute or arbitrary? Zero on the Centigrade Scale? Zero on the Avoirdupois Scale?

7. If a new pupil in your school made the following scores on Woody's Scales, Series B, to what grade do you think he should probably be assigned for arithmetic work? Addition, 14; Subtraction, 11; Multiplication, 12; and Division 6.

8. Suppose the pupil, whose scores were given in question 7, made only 45 as a score in the Buckingham Scale of Problems. What type of educational treatment would seem to be desirable? How much more useful would this type of information be to the new teacher than would the report that he had made a grade of C in fifth-grade arithmetic last year?

9. Assume that a pupil has a score of 3 on the Stone Test and a score of 70 on the Buckingham Scale. Which score would seem to you, considering the length of the tests and other similar factors, to be more worthy of dependence in entering the pupil in an arithmetic class in a new school?

10. Find the first, second, and third quartiles in the following distribution of scores on the Buckingham Arithmetic Scale.

Scores	40–44	45–49	50–54	55–59	60–64	65–69	70–74	75–79	Total
Frequency	1	3	8	13	12	7	3	1	48

REFERENCES FOR READING

J. Freeman Gay, et al., *Arithmetic Test*, Pittsburgh Public Schools, Research and Measurement Bulletin No. 2.

E. K. Sexton, *Arithmetic Survey in the Public Schools of Newark, N. J.*, Newark Board of Education.

C. W. Stone, *Arithmetical Abilities and Some Factors Determining Them*, Teachers College Bureau of Publications.

G. M. Wilson, *A Survey of the Social and Business Usage of Arithmetic*, Teachers College Bureau of Publications.

C. Woody, *Measurements of Some Achievements in Arithmetic*, Teachers College Bureau of Publications.

CHAPTER XIV

CLASSIFICATIONS OF MENTAL TESTS

INTERPRETATION of the results of mental measurements is less difficult when the scales by which the results were obtained have been carefully classified with respect to several important features. It must be recognized, however, that all classifications are made by drawing more or less arbitrary lines of distinction, which may not be equally applicable or useful in all cases. There is, for example, no definite point in the process of standardizing educational scales, below which a given scale is to be spoken of as "unstandardized" and above which it may be considered "standard." The term "standard scale" is nevertheless useful, if every one understands that it refers to a scale which has been relatively well standardized, in a relatively large number of respects. The distinctions and classifications made in the following pages should likewise be considered as relative and practical rather than as absolute and logical.

It is the purpose of this chapter to give useful points of view for examining and understanding the characteristics of educational measuring instruments, rather than to supply ready-made criticisms and classifications of specific scales. The writer hopes that one who reads the following discussion may be helped thereby in becoming more accurate and confident in his own judgments, not only with regard to

the measuring instruments now available but also with regard to others that will be published in the future.

Tests *vs.* scales. — Even the distinction between tests and scales mentioned in Chapter I is not always so clear and definite that one can be certain where the scale ends and where the test begins. Although the printed problems in the Courtis tests are the scale on which the child's performance is to be measured, it is not quite so clear that the time scale, on which the minutes are to be measured, should be considered as part of the test rather than as part of the arithmetic scale. When a boy's height is measured and found to be " 5 feet, $8\frac{1}{2}$ inches, in his stockinged feet," it is not always clear to every one that the phrase " in his stockinged feet " is a part of the test rather than a part of the measuring scale. It is nevertheless worth while to make the distinction between the scale and the conditions under which it is applied, and to discuss the relationship between the scale and its accompanying test.

It must be evident to any one who stops to consider the matter that a scale by itself is not a particularly useful instrument. A native in the wilds of Africa might find no use whatever for a clinical thermometer, although such a scale is often extremely valuable to the physician, or to any one who understands the rules for using it in testing the body temperature of an invalid. Similarly, the scales of the Cleveland Survey Arithmetic series are not especially significant or helpful, except in the hands of some one who knows the rules of the test. On the other hand, no test is of great value if the scale applied is inaccurate, or — what is still worse — if the scale appears to be much more accurate than it really is. Valid measures or scores in any

field can be obtained only by the employment of the most accurate scales under the most carefully defined test conditions. Tests and scales can be useful only when employed together, and to obtain the most useful measurements both the tests and the scales must be the best obtainable for the purpose.

Objective *vs.* subjective scales. — It is highly desirable that any scale proposed for general use be examined critically to determine the extent to which it is objective rather than subjective in character, since no scale for measuring mental facts is free from subjective factors. As was pointed out in the first chapter of this volume, there is a slight subjective or personal variation even in the readings made by different scientists of the same standard yardstick. If perfect agreement is not possible in the physical field one need not be surprised or discouraged at the discovery of even greater subjective elements in the field of mental measurements.

One factor which contributes much to the objectivity of any test result is permanency of record. If the result of the child's mental activity disappears as soon as the activity is ended, one may expect a relatively low degree of objectivity in its measurement. Few attempts have thus far been made to measure oral composition, for example, because the making of short-hand or dictaphone records of the child's oral composition is expensive and distracts the child's thought from his work. Unless competent judges, who were not present when the original mental act was being performed, may examine the evidences of the action and compare their judgments of them with the judgments made by the judges who were present, subjective elements are certain to play a dominant part in the ratings given.

It is for this reason that "pencil and paper" tests have become almost universal. It is desirable to measure what children think, or feel, or do, or say, under many different types of situations, but it is extremely difficult to obtain permanent and accurate records of many types of thoughts and feelings. It is, on the other hand, fairly simple to obtain a permanent record of what a child writes when a printed question is presented to him. Unfortunately, one cannot be sure that, under given conditions, a child or an adult would actually think, feel, or act in just the manner he believes and writes that he would. Human beings enjoy the approval of their fellows so much that most of us, if a given situation is described to us, are inclined to say we would do or think or feel just about as we have been taught good people ought to behave in such circumstances. In a paper and pencil test, no intelligent person would admit that he believes in ghosts, and yet some who would not admit it on paper show signs of real fear when going through a cemetery alone on a dark night.

One important measure of the excellence of a scale is therefore the degree of its objectivity, the degree to which an amount expressed in its terms can be identified, examined, and verified by competent persons. Different people have had different experiences and have come to assign slightly different adjectives to the same amounts of quality, so that one person's "excellent" may mean exactly the same amount of quality as another person's "fair." Permanently preserved samples of various degrees of goodness must be available for the study of all who are interested, before one can claim great objectivity for any mental or social scale.

Mental *vs.* physical scales. — Every mental experience or activity is accompanied or occasioned by physical changes in the body of the person to whom the experience belongs. These changes may be little more than the transmission of impulses over certain nerve fibers, or they may involve the activity of many systems of facial or bodily muscles. Whatever the extent of the bodily changes, they are always present. One must recognize that no mental operation is wholly independent of physical effects. On the other hand, it is probably just as true that no physical change occurs without some slight change in the mental condition of the individual concerned.

It is therefore impossible to declare confidently that any mental test is free from the effects of physical conditions, or to assert that any physical test is wholly independent of mental controls. It is useful, however, to distinguish certain tests as primarily mental and others as primarily physical in character. To make this distinction with full confidence is most difficult in the cases of tests for measuring the rate or speed at which one can work. It is not always clear in such tests whether the score is an index of rate in mental operations or only of the rate at which the answers can be written. Every mental test should be studied before it is used, to determine the probable influence physical limitations have upon it.

Educational *vs.* intellectual scales. — It is particularly difficult to make and to use the distinction between tests that measure primarily the results of original or native intelligence, and tests that measure chiefly the results of instruction and environment. One cannot learn unless one has an adequate supply of appropriate intellectual

fibers to be changed and improved. On the other hand, intellectual growth is undoubtedly stimulated as well as demonstrated by the improvements or changes resulting from instruction. There are psychologists who think of intellectual ability as being a summation of all the mental powers of the individual, — an epitome of the original nature of the individual plus all the improvements and changes wrought in his nervous system by school training as well as by experiences outside the classroom. From this point of view, intelligence is a general term that includes all the special mental abilities, regardless of whether they were developed at school or elsewhere.

The tests now used for measuring intelligence are, for the most part, tests of a mixture of many different abilities rather than of any one specific ability. It seems impossible to get a test which will give just as high scores to the " intelligent but untutored " man as it gives to the man of equal intelligence who has had the benefit of good training. As a matter of fact, it is difficult to find a sufficient number of persons of high intelligence, who have not also had educational training, to prove that any particular test is really independent of education. Our so-called intelligence tests are not radically different therefore from the educational tests, except that they sample many different kinds of mental ability and mix the results into a general score.

One should not receive the impression, however, that the *only* difference between an educational test and an intelligence test is that one is specific and the other general and comprehensive. Certain types of educational measurements, particularly measurements of the rate at which pupils do equally difficult tasks, are relatively poor measures

of general intelligence. It must also be said that the contents of the intelligence tests are usually chosen from fields of information and experience not directly supplied by the ordinary course of study in the public schools. Most of the facts called for in an intelligence test would be readily "picked up" by any bright child who merely grew up in a community of intelligent people, even though the child himself did not attend school. But reading the newspapers, bill-boards, and school textbooks would certainly give a pupil some advantage in almost any intelligence test now available.

The reader should be especially alert in studying certain paired groups of tests now being sold in a single booklet, one group being called educational and the other intellectual. It would probably be more proper and exact to call one group of these tests an academic intelligence test and the other group a non-academic intelligence test, for both groups of tests yield rather general measures of various phases of intellectual alertness. A test should not be classified as being purely an educational test just because it involves reading and other skills taught in schools, nor should a test be finally classified as an intelligence test because it does not involve reading and other formal school subjects. As was stated above, the distinction between educational tests and intelligence tests is difficult to make and still more difficult to use. It would often be better to make no such distinction at all than to reach false conclusions because an arbitrary distinction had been made.

Reading *vs.* spelling, writing, and arithmetic. — It is almost impossible to develop a test in one school subject that does not measure some elements of other school sub-

jects. The ability to read and to interpret the language in which problems in arithmetic are stated is included in every arithmetic reasoning test. Similarly, the ability to read is involved in geography tests, history tests, science tests, writing tests, and in the tests of almost every other special field. Writing ability is likewise implied in most of the tests of ability in special fields. Ability in elementary mathematics is frequently called for in some of the so-called reading tests. In fact, a test of reading ability alone, or of arithmetic ability wholly apart from and uninfluenced by other abilities, should probably not be expected. One should, however, be quite conscious of the extent to which his scores in a test having a given name are modified by other types of ability. He should also seek, in so far as possible, to use in each field a test that measures primarily the ability which it purports to measure.

Group *vs.* individual tests. — The distinction between group tests and individual tests is one of the easiest to make. A test which is to be given to one pupil or person at a time is called an individual test, while a test that may be given to several people at the same time is called a group test. It is possible, of course, for almost any group test to be used as an individual test. It is not possible, however, to use many of the individual tests as group tests. The records of an individual examination are usually made by the examiner rather than by the one examined. The one who is giving a test cannot conveniently observe and record the simultaneous responses of several different persons, or if he attempts to do so he will often find his subjects imitating each other rather than giving their own responses. Group tests are only possible where each individual's re-

sponses can be permanently recorded by himself without disturbing or influencing the responses of other individuals.

The advantages of the group tests are that the responses are definite and can be reëxamined as often and by as many different judges as desired, and that they are economical in respect to the examiner's time. There are few advantages in the individual examination except when given by an expert who has had long training and experience. For such an expert, the manner of a person's response to a question is often much more significant than the response itself. It may surpass the group test in completeness. Those who are not well trained or are without the clinical experience should not usually attempt to use the individual tests. It is more economical for the average teacher or supervisor to use several well-selected group tests than to attempt to give an individual examination to each pupil.

Children's scales *vs.* adult's scales. — Just as there is no sharp line between youth and age, so there is no clear-cut line to be drawn between scales for measuring children and scales for measuring adults. In fact the ideal scale would be one which would measure all degrees of any given mental trait, just as the avoirdupois scale measures all degrees of weight or as the linear scale measures all degrees of length. There has been a tendency to fear that children, who were measured on a mental scale whose higher limits were far beyond their reach, would be discouraged thereby, or possibly that they would injure their minds by trying to reach the higher limits. As a result there are many elementary grade scales, intermediate grade scales, junior high school scales, senior high school scales, and college scales.

This specialization of scales for persons of each age or de-

gree of maturity is unfortunate in many respects. First of all, it removes the possibility of comparing directly the performances of people at different maturity levels. The facts regarding the overlapping of the mental abilities of one group on those of a more advanced group are always startling and instructive, but such facts are not easily shown when a different scale is used for each group. It is altogether proper to break up an entire scale into sections, as in the case of Buckingham's Scale for Problems in Arithmetic, discussed in a previous chapter; but the final measurements should be in terms of the original scale regardless of the special section of it used on a particular group. Unless the same scale is used in measuring all groups, such facts as those represented in Table XXII and Figure 26 of Chapter VII will not be discovered.

In the second place, increasing the number of scales to be used in measuring any mental trait tends to increase the confusion in the minds of both laymen and teachers as to what the results mean. Unless tests are given for the purpose of answering definite questions, they are often worse than useless. If each segment of the school system is to be measured in a given subject by means of a different scale, having in many cases no known relation to the scales used in measuring the same subject in other segments of the school, many of the problems that measurements should solve will be no nearer solution than they were before tests came into existence. Where it is at all possible, each educational scale should be capable of measuring the ability of both pupil and teacher. The science of education will be built up most rapidly when tests are devised to measure the results of instruction in a given field on a continuous

scale, from the kindergarten to the university. Educators may have to be satisfied temporarily with one set of scales for the primary grades and another set for the higher grades, but they should make it possible ultimately to measure all grades with a single set of continuous scales.

Universal or continuous scales cannot be constructed as quickly and carelessly as many of the shorter scales are now constructed. There is great danger that a scale of extensive range will be inadequate at the extremes. One of the most difficult tasks in scale construction is to secure an adequate supply of extremely high elements and extremely low elements. Unless the scale contains some elements that are beyond the range of the best pupil, that pupil will remain unmeasured. On the other hand, if the lowest element in the scale is not easy for the lowest pupil, that pupil cannot be measured. Even when the scale extends far enough in each direction to include all the extreme cases, there is the danger that the steps near the ends of the scale may not be equal to the steps in the middle. Such unequal steps would distort the facts concerning the relative positions of the pupils at the extremes when compared with average pupils.

Measuring processes *vs.* measuring results. — Some teachers may feel that the distinction between products and processes is not necessary. The aim of the school should always be to help the pupils to master certain processes, and the obvious way to discover whether pupils have mastered a given process is to examine the results they produce. If a pupil gets the wrong answer to a problem, there must be a flaw in his work. If the cake " falls," then the girl must have failed to master completely the process of baking cakes.

If the boy has not done what the printed directions in a reading test tell him to do, he must have erred in his reading of the directions. It is difficult to think of a school subject in which the mastery of an operation is not the important consideration, and in which the result of the operation is not the basis on which one may judge whether or not the pupil has mastered it.

In spite of this general relationship between school products and the processes by which they are produced, one may not make valid comparisons between the efficiency of two schools on the basis of their products alone. An exhibit of five dresses made by the girls of School A might be superior to an exhibit of five dresses made by the girls of School B. This fact in itself would not prove, however, that School A was superior to School B in dressmaking instruction. School B may have spent only three recitation periods in preparing their exhibit, while school A may have spent several months. There may have been only five girls studying dressmaking in the second school, while the first school may have selected the best of the products made by a hundred different students. The girls may have done all the work themselves in school B, while the teachers may have selected the materials and patterns in School A. A knowledge of the processes by which the dresses were made will help to determine which of the two schools teaches dressmaking more effectively.

The operations by which the results were obtained must be known just as thoroughly as the results themselves whenever one undertakes to make comparisons as to efficiency. Two boys may each have won ten games of tennis on a certain day, but one may have won against a beginner

while the other may have played against an expert. A friend once boasted to the writer that the horse he was driving " won third money " in a race at the county fair. Inquiry brought forth the fact that there had been only two other horses in that race. Two boys may each be known as " the brightest boy in school," but one of them may be in a very small elementary school while the other may be in a large city high school with two thousand other bright boys. The conditions of a contest must be equalized; otherwise a result has no meaning.

Sometimes a fair contest emphasizes the results produced under equalized conditions, while at other times the emphasis is on the conditions under which equal results are produced. The horse race is a good illustration of a contest in which the action varies while the product is the same for all contestants — a run of one mile. It would be perfectly possible, of course, to compare two horses on the basis of their products — the distance traveled in five minutes' time on a level track. The custom has grown up, however, of measuring the process rather more directly than the products of such races. The old time " ciphering match " was of the same character as the horse race. Each contestant was given the same problem to solve, but that person who solved it correctly and most quickly was declared the winner. The modern Courtis tests, described in a previous chapter, have reversed the emphasis. Instead of measuring the operation by which a uniform result is produced, the Courtis tests measure the amount produced under uniform conditions — the number of equally difficult addition problems correctly solved in eight minutes.

The writer does not intend to advocate the measurement of educational work by any one of these schemes to the exclusion of all others, but he does wish to emphasize the necessity of knowing the nature of the scheme used in any standard test. One should never use any test in a school until one has satisfied oneself on the following questions: What will be the chief element of difference between pupils revealed by their scores in this test? How are the other variable elements in the test related to this dominant element? Until these two questions have been satisfactorily answered, one will not be able to interpret the true significance of the results of any educational measurement.

Amount *vs.* quality of educational products. — When the time and difficulty of the process have been made equal for all contestants, two important elements about a product must be measured accurately: its amount and its quality. One must know how many pies were baked and how good they were; how many problems were solved and how nearly perfect the solutions were; how many letters were written and how well they were written; how many words were spelled and how perfectly they were spelled; how much sewing was done and how well it was done.

Certain of the best educational scales were intended to do no more than to measure the quality of the products prepared by the children, regardless of the size of these products or of the process by which they were obtained. Other scales are primarily counting devices for determining the amount of the result obtained. Unless the various characteristics of process and product are related to each other in a definitely known degree, the conditions of the test must allow only one characteristic to vary at a time. Otherwise

it would be impossible to obtain a relative efficiency rating for two or more persons. For example, there is no possibility of saying with confidence that the pupil who wrote 65 letters of quality 10 in one minute did better or worse than the pupil who wrote 70 letters of quality 9 in one minute. Similarly, no one knows whether the pupil who in twenty minutes draws pictures of ten men, each picture being of quality 12, is a better or a poorer artist than the pupil who in twenty minutes draws pictures of two men, each picture being of quality 16. Here, in order to compare the artistic efficiency of the two pupils, it would be necessary to measure either the quality of a picture each can draw in the same length of time, or the number of pictures of the same identical quality each can draw in a given time. To allow both the quality and quantity of a product to vary at the same time removes the possibility of making accurate comparisons directly.

The first modern attempt to prepare an educational scale was an attempt to measure the quality of a school product. Thorndike published in 1910 a *Scale for Handwriting of Children in Grades 5 to 8*.[1] This scale makes it possible to measure the general quality of the child's handwriting, regardless of its amount or of the conditions under which it was produced. The various degrees of quality illustrated by the samples on the scale were determined by the combined judgments of persons whom Professor Thorndike considered competent to judge handwriting. When one uses the scale, he merely tries to find on the scale a sample which has approximately the same amount of general quality, " goodness," or " merit," as the sample being measured.

[1] E. L. Thorndike, "Handwriting," *Teachers College Record*, Vol. II.

It would be possible, of course, to obtain scores for the quality of a product in other ways than by having it rated on a scale composed of increasingly good samples. The amount of disagreement between different teachers might be reduced from what it usually is, if all teachers were to agree that a certain sample was worth 100%, that a certain other sample was worth 0%, and that all grades for quality were to indicate whether the sample being rated was nearer to 0 or to 100, and (as accurately as possible) how much nearer. In rating the qualities of most objects, teachers have quite different 0 points and quite different 100% points. It is little wonder that they disagree in their ratings.

Still another method of obtaining relatively objective measures of the amount of quality in an object is illustrated by the score-cards in use by agricultural colleges for measuring the quality of an ear of corn or the merit of a hen. This method calls attention to all the good points of a perfect product and gives each of these points a certain weight, depending upon its importance in the total amount allowed for an example of perfect quality. The superiority of such a method of rating over the usual uncontrolled and undefined rating is that it tends to limit each judge to a common maximum on each item. Such a score-card in handwriting, for example, would prevent a judge who was an extremist from deducting half of the total possible number of points just because the pupil had not maintained a uniform slant in his letters. It would prevent some other judge whose hobby was " spacing " from deducting more than should be deducted for a faulty spacing of the letters and words. The score-card method of measuring the quality of a product not only describes a sample of perfect quality but it

also limits and guides one who uses it, making his ratings conform more closely than they otherwise would to the ratings other judges would give for the same object.

Even the score-card method allows a considerable amount of difference between the ratings of different judges. If a score-card could be illustrated by graduated samples, the differences between judges might be still further reduced. It would probably be better, for example, if there were in the Strayer-Engelhardt Score-Card for City School Buildings a series of six or seven photographs, each showing how many points it deserved out of the total of 35 points for perfect "Size" and "Form." With such objective aids, it would perhaps be much easier than at present for a superintendent, making use of the score-card, to arrive at a total score which would agree closely with the score Professor Strayer would give the building. Such a combination of the score-card method and the scale-of-increasingly-good-samples method seems to be the direction in which the measurement of quality in educational products will probably develop in the future. The graduated scale for general quality often represents distinctions which an average judge does not even see, until his attention is especially called to them, while the score-card frequently leaves the user at a loss to know how much to deduct from a perfect score on account of a partial defect.

The Courtis Arithmetic Tests and the Cleveland Survey Arithmetic Tests are excellent examples of scales that measure primarily the amount or size of a product. One must be very careful in such scales to make certain that the units are equal. It is so easy to believe that "spelling three words correctly" is just as much less than "spelling

four words correctly" as spelling four words is less than spelling five. The equality of units of amount must be well demonstrated before one can draw sound conclusions as to relative efficiency on the basis of measured amounts of product, even when its quality and the conditions under which it was produced have been equalized.

Time *vs.* difficulty of producing a result. — Two factories may be producing exactly the same amounts and qualities of results without being equal in effectiveness. One may produce the specified amount and quality of output by running eight hours per day while the other may run twelve hours per day. Or the one may turn out the given quality and quantity of product by combining prepared materials, while the other may carry through all the operations, beginning with the raw materials. It is not safe to judge two factories as equal in efficiency just because they produce the same number of yards of equally good cotton cloth or the same number of equally good automobiles. The time consumed and the difficulties overcome in bringing forth a product are just as important in determining success as are the amount and quality of the results.

The time required to accomplish a given result may be easily measured by any accurate watch, clock, or other chronometer. It is important, however, to know exactly when to begin counting the time and when to stop. Unless the beginning of the process and its completion are counted in the same manner for all competitors, the records of time required will not give an accurate index of relative success. In measuring educational results, one must observe carefully the rules for measuring the time during which the pupils are at work.

Measuring the amount of difficulty overcome in carrying any operation through to completion is a complex task, and yet it is one of the most useful types of measurement that one can make in educational work. It is much easier to prepare tests which measure exactly how many equally difficult tasks one can do perfectly in a given time than it is to prepare tests that measure accurately how hard a task one can do perfectly in a given time. There are many purposes, nevertheless, for which the tests measuring the power to meet increasingly difficult situations are more helpful to the teacher or supervisor than are tests measuring the rate at which pupils do equally difficult tasks.

In general, the most useful basis for determining which of two tasks is the easier is to find out which can be done successfully by the larger number of persons. If 90% of all eighth-grade pupils in a city can solve problem A correctly, while only 50% of these same pupils can solve problem B correctly, one may assume that problem B is more difficult, for these eighth-grade pupils at least, than is problem A. Other means of measuring relative difficulty have been proposed in some cases, such as the relative length of time required to solve the two problems, but no other method has proved quite so satisfactory and convincing as that of the relative number of persons able to solve each problem. Estimates of experienced teachers may sometimes be used as preliminary bases for ranking questions as to difficulty, but the real proof is found in the relative success of pupils in answering them.

In tests measuring primarily the difficulty of the work pupils can do, the time allowed for the work and the amount and quality of the results should be kept uniform for all

contestants, except where one or more of these elements can be subordinated to the dominant element of difficulty. If the amount of the product were unrelated to the difficulty of the work done, for example, and if both were allowed to vary during the same test, the competitors with the largest amount of product might conceivably have overcome less difficulty than those whose amount of product was smaller. In actual practice, the questions or problems of least difficulty are usually placed first and those of most difficulty placed last in such a test. The pupils are then required to begin with the easy examples and to answer them in order. The time limits, in standard tests for measuring difficulty, are so generous that a child would hardly be able to answer a significant number of more difficult questions if allowed more time.[1]

The scoring of tests measuring difficulty of performance. — Tests in which the difficulty of the performance is the dominant variable have suffered severely at times in re-

[1] Quite a different attitude on this matter is expressed in Chapter VIII of *The Measurement of Silent Reading*, by Dr. May Ayres Burgess. She states that "The scores which result from such testing represent genuine difficulty scores for some children, rate scores for other children, and varying combinations of difficulty, rate, and quality, for still others. They are not comparable, because they do not measure achievements of a single kind. They are conglomerate scores of difficulty, quality, and amount, in unknown and varying proportions. They do not conform to the principle of the single variable."

For evidence that difficulty is the dominant element measured and that the time allowance has little effect on the scores in standardized difficulty tests, see Table LXX on page 294 of *Measurement of Classroom Products*, by Stuart A. Courtis, New York General Education Board.

spect to the method of assigning scores for work done. The plan proposed by Thorndike [1] in connection with his reading scales, whereby a class score was determined on the basis of the percentages of right responses to those elements of the test causing nearest to 80% of correct answers, was never well understood or widely used by teachers. The general educational public likewise failed to grasp and employ the modification of Thorndike's plan proposed by Woody at the suggestion of the present writer.[2] The yet more comprehensive and thorough plan proposed by Truman L. Kelley [3] received still less general acceptance. The above schemes aimed to obtain reliable scores in terms of how difficult a task could be done, but the general public seemed to want easily obtained scores for the individual pupils, regardless of how low their reliability might be.

The most practicable plan yet devised for assigning difficulty scores to pupils seems to the writer to be that of Buckingham in his Scale for Problems in Arithmetic. In this scale, a child's score is based on the difficulty of the last (most difficult) problem solved correctly. The increments of difficulty between consecutive problems in this scale are about three points. Three points are therefore deducted from each pupil's basic score for each easier problem incorrectly solved. This method is so simple that most teachers can use it, and yet it gives a good statement of the

[1] E. L. Thorndike, "The Measurement of Ability to Read," *Teachers College Record*, Sept., 1914.

[2] C. Woody, *Measurements of Some Achievements in Arithmetic*, Teachers College, Columbia University.

[3] T. L. Kelley, "Individual Testing with Completion-Test Exercises," *Teachers College Record*, Sept., 1917.

pupil's approximate location on a scale for mastering difficulties. It is to be hoped that this method will meet with general favor and acceptance.

The most widely used scheme for assigning scores, on scales which measure difficulty of performance as the dominant variable, is the extremely simple one of counting how many items were solved correctly. This method does not greatly distort the facts when there are equivalent intervals between consecutive problems, as represented in the line of x's given below. The child who solves the first five

Easy	Hard
x x x x x x x x x x x x x x x x x x x x x x x x x x x x x	

problems correctly will receive a score of 5; the one who solves the first ten problems will receive a score of 10; and the one who solves fifteen will get a score of 15. The difference between 10 and 15 will appear and will actually be about equal to the difference between 5 and 10, and about half the difference between 5 and 15.

This scheme distorts the facts very greatly, however, when the spacing between the difficulties of consecutive problems

Easy						Hard
xx	xxx	x		x	xx	xxxxxxxx

is not uniform. It may well happen that the actual difference between doing 5 and doing 10 will be four or five times as great in terms of difficulties overcome as the difference between doing 10 and 15. If the score for difficulty over-

come is to represent honestly the differences between pupils, the differences in difficulty between consecutive examples must be uniform.

One of the most stupid practices in scoring consists in adding, for each additional increase in difficulty, the sum of that increase plus all that had previously been accomplished. If one child has a score of 10, for example, and another child answers all the same questions plus another that is 1 point more difficult, this absurd method of scoring would add to the pupil's score, not 1 point, but 11 points. This is on exactly the same principle as though one were to weigh a baby and find that he weighed twenty pounds; to weigh him a month later and find that he had gained a pound; and then to declare that the baby's weight was forty-one pounds ($20 + 21 = 41$). This seems absurd, and yet some widely used educational tests are scored on exactly this principle.

Tests of rate *vs.* tests of power. — By rate tests or speed tests are meant tests in which the dominant variable is quantity of product produced. In such tests the other three variables, quality of product, time of production, and difficulty of production are kept constant, while the number of products varies according to the abilities of the contestants. The Courtis Tests, the Cleveland Survey Arithmetic Tests, and the test for speed of handwriting are examples of rate tests. The scores in rate tests are easily influenced by the recency and frequency of drill work.

By power tests one usually means to indicate tests in which the dominant variable is the degree of difficulty overcome in doing assigned tasks. In such tests, the quality of the product remains constant (for no answer is counted

unless it is perfectly correct), the amount of the product is so arranged as to indicate the degree of difficulty successfully met, and the time is uniform for all pupils and sufficiently long that all will have reached their limits before being stopped. The Thorndike reading scales and the Buckingham Scale for Problems in Arithmetic are good examples of the so-called power tests. More fundamental instruction and explanations are required in order to effect changes in the scores on power tests than are required to change the scores on rate tests. The power tests agree more closely with intelligence tests than do rate tests, which fact has caused some writers to express the belief that power tests should be considered general ability tests rather than tests of ability in educational subjects. The writer considers the power tests just as truly tests of ability in specific subjects as are the rate tests.

Summary. — This chapter classifies measurements according to several criteria, some of which are arbitrary rather than logical. The purpose of these classifications is to call the attention of the reader to points of view from which specific tests and scales may be studied critically. Of special importance is the distinction between objective and subjective tests, and the attempt to distinguish intelligence tests from educational measurements. The doctrine of the dominant variable is advanced, and four variable elements are given places of honor: time of the activity, difficulty of the activity, quality of the result, and amount of the result. Relative efficiency may not be determined accurately except in terms of one variable at a time. It may sometime be possible to evaluate correctly the units of one variable in terms of another, but at present

a test is not fully satisfactory if its scores are indicative of more than one of these variables.

QUESTIONS AND EXERCISES

1. Under what circumstances might it be advisable to accept the results of subjective ratings rather than of objective tests?

2. What human characteristics will probably be the last to yield to measurement by "pencil and paper" tests?

3. What types of use can one make of metrical or English measurements of distances that could not be made if short distances were measurable only by the Turkish *archine*, slightly longer distances measurable only by the Chinese *pu*, and still longer distances only by the Russian *verst?*

4. It has been suggested that the amount of credit to be allowed for solving a given problem should be determined by the frequency with which that problem will be needed in later work rather than by the frequency with which pupils fail to solve it. What validity has this suggestion?

5. The Rugg-Clark Algebra Tests are primarily rate tests, while the Hotz Algebra Scales measure difficulty of problems solved. Which series would you expect to show the closest agreement with other tests of education and intelligence?

6. In what sense may educational measurements be said to be more concerned with products of pupils than with processes in pupils?

REFERENCES FOR READING

M. A. Burgess, *The Measurement of Silent Reading*, Chapters IV to IX.

Harris and Breed, "Comparative Validity of the Hotz Scales and the Rugg-Clark Tests in Algebra," *Journal of Educational Research*, Vol. VI, pp. 393–411.

W. S. Monroe, *Theory of Educational Measurements*, Chapter IV.

CHAPTER XV

MEASURING ABILITY TO READ

THOROUGH agreement as to the meaning of the term "ability to read" is hardly to be expected. The census for the year 1910 reported that approximately 8% of the population of the United States was illiterate, while the psychologists who examined the men drafted for military service in the United States Army reported in 1918 that about 25% of these men were illiterate. This wide difference in the percentages was not because men between the ages of twenty-one and thirty-one were less literate than the general population, but because ability to read and write as defined by the psychologists was different from that ability as defined by the census-takers. On account of such differences in definitions, those who are to-day undertaking to measure the amount of reading ability possessed by pupils of the public schools differ in their results almost as much as did the psychologists and the census-takers.

It is not surprising that there is a lively debate among those who are most expert in educational measurements as to what constitutes ability to read and write. So long as people discuss matters in vague general terms, with the unsupported assumption that every one understands what is being said, it is relatively easy to have harmony and perfect agreement. But as soon as one undertakes to define the terms of the discussion in exact, objective terms, vital

differences in points of view are quickly discovered. The general public was not much concerned over the fact that almost 8% of the population was "illiterate," and no one felt called upon to question the fact or to argue about it. A warm discussion immediately arose, however, when it was reported that one of every four adult men was "unable to read well enough to understand an ordinary newspaper and to write well enough to communicate with his friends." An exact statement of facts can be questioned and verified, while a vague, general statement may mean almost anything its author wishes it to mean. The lack of agreement among educational measurers is therefore a hopeful sign pointing toward the early development of a real science of reading.

Possible units of ability to read. — Reading ability is by no means a simple characteristic which can be quickly acquired by any one who desires it. The ability to read is one of the most complex traits mankind has developed, and it is probable that a fairly large percentage of the human race is incapable of a high degree of achievement in reading. It is such an important tool of civilized peoples, however, that its analysis and measurement are necessary if education is to make the maximum amount of progress.

The distinction between reading aloud and reading silently is one of the most obvious as well as one of the simplest distinctions to be made, and yet it is only within recent years that any careful attention has been paid to this difference. Whenever the question is raised as to how well a person can read, the first subordinate question to be answered is that which differentiates between these two reading abilities. Silent reading is more universally important

than oral reading, and will therefore be used to illustrate the general principles.

How may one compare the amounts of silent reading ability possessed by different pupils? If the comparison is to be perfectly just and accurate, the four variable elements mentioned in the previous chapter must be measured, one at a time. There must be a measurement either of the *time* required to read, at a given degree of excellence, material of a given amount and difficulty; or of the *excellence* of the reading of material of a given amount and difficulty during a fixed time; or of the *quantity* of equally difficult material read in a given time at a fixed degree of reading quality; or of the reading *difficulty* reached in a given time over a fixed amount of material without loss of quality in the reading. In other words, some one of these four variables must be dominant over the others in order to make a reading test worthy of recognition as a valid measuring instrument.

Units of quantity may be given first attention. Shall the *letters* of the alphabet be considered satisfactory units for measuring the amount read, or is it necessary to employ larger units? Perhaps it would be better to count the number of *words* read, or possibly the number of *phrases* or *sentences*. It is possible, of course, to build a reading test with *paragraphs*, *sections*, or *chapters* as the units of amount. It is hardly probable, although perhaps possible, that someone will make a scientific reading test that will employ the *volume* or *book* as the unit of measure.

The units for measuring the quality of reading — how well the reading was done — will depend in large measure on what unit has been chosen for measuring the amount.

If the amount done is to be determined by counting the number of "words read," for example, the excellence of the reading may be considered on several different levels. There has been at least one proposal to have the pupils check each word they recognize, which amounts practically to reading just well enough to indicate that the word has been seen before. It must be admitted that this is a low degree of reading quality, and that it would be easy for a dishonest pupil to gain an unfair advantage in such a test. There is general agreement that a thoroughly reliable test cannot be built on such a unit for measuring quality of work done.

A more useful proposal is that one read well enough to indicate to what general group of things each word refers. In the Thorndike Visual Vocabulary Scales, for example, the child is required to "write the letter F under every word that means a *flower*, the letter A under every word that means an *animal*," and so on. It would perhaps be possible to devise a test in which one should read each word well enough to give a useful definition or description of it, but so much of the time in such a test would be taken for the definitions that little would be left for the actual reading. Professor Thorndike has recently published an excellent Test of Word Knowledge in which pupils are required to read each word well enough to select its synonym from a list of five words and phrases.

An interesting attempt to devise a word reading test that would be useful in the lowest school grades has been made by Pressey and Skeel.[1] The proposal is to have the

[1] L. W. Pressey and H. V. Skeel, "A Group Test for Measuring Reading Vocabulary in the First Grade," *Elementary School Journal*, December, 1920.

pupils pick out and check those combinations of letters which are real words from among other combinations which are nonsense, in such lists as the following:

tl mz the unf pr
wb I su en w
ob buq bl lz be
mt tu en is de

Whether first-grade pupils have sufficient familiarity with word forms to furnish in this manner a reliable record of their abilities is doubtful. The use of this level of quality in reading (sufficient understanding to check the real word in a group of nonsense syllables) is, however, an interesting attempt to find a satisfactory standard of quality for immature pupils.

Similar variations in units or standards of reading quality might be found for each different unit of reading quantity. If the amount read is measured by paragraphs, the quality or character of the reading might be standardized at " just well enough to recognize that it says something," or at " just well enough to understand what it says," or at " just well enough to be able to do what it says to do." As in the case of reading individual words, however, there is the possibility at the first two levels that a dishonest child might cheat. This necessity of eliminating as nearly as possible the chance for unfairness, requires that the unit of reading quality shall be sufficiently high and complex that the examiner may know, from the recorded actions of the pupil, whether or not he really read the word, the sentence, or the paragraph. Most of the reading tests therefore contain directions to the pupil to make a distinctive pencil mark on or near the words and pictures before him.

The most frequent requirement of the pupil is that he answer certain questions about the material he has read. In such cases the questions are, of course, just as much a part of the scale as are the sentences, paragraphs, or other units of material. The child's written answers to the questions give permanent, objective evidence of the quality of his reading. In many cases the child is not required to write answers but only to check the correct answer from a list of several printed suggestions. The checking of a correct answer from a printed list requires less time and does not demand as complete an understanding of the materials as is required where the pupil must find the answer without suggestions. Some writers hold that the high degree of selective thinking, which is necessary in order to devise and to write the answer to a question, should not be required in a reading test. It is desirable, certainly, that the teacher be aware of the degree to which various other abilities are included in any reading test.

The selection of a unit for measuring the difficulty of a word or of a paragraph is almost altogether a statistical problem. The methods involved will be discussed at some length in Chapter XX, but it is worth while to state here that the only satisfactory basis yet found for measuring difficulty is the relative frequency with which a given task can be done. A task which every one can do is either more familiar or less difficult than a task which can be done by only half of the population, and difficulties in mental operations are chiefly lack of familiarity with the processes. One of the most widely used reading scales, however, attempts to measure relative difficulty in terms of the relative amount of time required to accomplish each task. It is quite evi-

dent, to one who thinks carefully about the matter, that task A might be so easy that all pupils could do it if given twenty minutes' time, while task B might actually require only two minutes' time but be so difficult that only one out of ten people could ever do it. Until some better basis is found, estimates of the difficulty of a task should be based upon the theorem that " tasks which an equal number of persons can perform are for that group of persons equal in difficulty."

Having considered the various possible units for measuring quantity, quality, and difficulty in reading, the reader may now turn his attention to a few of the tests that have actually been prepared and used. It will be well to note in each test just what units or standards are employed for measuring or controlling each variable.

Gray's Oral Reading Test. — The most important attempt to measure results in oral reading have been made by Dean William S. Gray of the Chicago University School of Education.[1] The material of Dean Gray's test consists of twelve paragraphs, each one of which is somewhat more difficult than the one before it. The first three paragraphs are printed in large type such as is frequently used in children's books. The first, fourth, eighth, and twelfth paragraphs are reproduced below to show the character of the materials.

[1] See pages 263–275 in the volume by C. H. Judd entitled "Measuring the Work of the Public Schools," in the *Cleveland Foundation School Survey Report*.

Since this discussion was put into type the Public School Publishing Company of Bloomington, Ill., has published a new Oral Reading Check Test by Dean Gray.

SELECTED SAMPLES FROM GRAY'S ORAL READING TESTS

1

A boy had a dog.
The dog ran into the woods.
The boy ran after the dog.
He wanted the dog to go home.
But the dog would not go home.
The little boy said,
"I cannot go home without my dog."
Then the boy began to cry.

4

Once there lived a king and queen in a large palace. But the king and queen were not happy. There were no little children in the house or garden. One day they found a poor little boy and girl at their door. They took them into the beautiful palace and made them their own. The king and queen were then happy.

8

The crown and glory of a useful life is character. It is the noblest possession of man. It forms a rank in itself, an estate in the general good will, dignifying every station and exalting every position in society. It exercises a greater power than wealth, and is a valuable means of securing honor.

12

The hypotheses concerning physical phenomena formulated by the early philosophers proved to be inconsistent and in general not universally applicable. Before relatively accurate principles could be established, physicists, mathematicians, and statisticians had to combine forces and work arduously.

Each child examined must be taken individually by the examiner to a quiet place, where the reading will not interfere with the work of other pupils or be interrupted by other activities. As the child reads aloud the examiner keeps a careful record of the time required for each paragraph and of the mistakes made, such as omissions, repetitions, and mispronunciations. The score made by the pupil depends not only on his rate of reading and its quality, but also on the grade or class from which he comes. A fairly complete set of blanks and directions is supplied by the author for keeping the records and making the calculations.

The Gray Oral Reading Test is the best available at the present time, but it has many unsatisfactory characteristics, some of which cannot be avoided in the measurement of an oral performance. The most fundamental difficulty with a measurement of oral work is that the child's reactions are not automatically recorded where any competent person may later judge of their quality. Unless an examiner has been carefully trained and tested by the author of the test, his record of a child's performance will probably differ considerably from that which Dean Gray would make. The use of a dictaphone or other recording instrument would probably be somewhat distracting to the pupil and

too expensive for the examiner, although it is probable that before many years oral tests will become much more exact and objective through the use of some such automatic recording instrument. The judgment of the examiner plays too large a part in determining the score on oral tests at present.

It is unfortunate that the method of scoring suggested by Gray makes it impossible to compare the scores of pupils in one grade directly with those of pupils in higher and lower grades. The analysis of results by ages within each grade is also less helpful than it would have been if the scoring were the same for all pupils. The fact that only one pupil may be tested at a time makes the test expensive and has probably been the largest factor in preventing it from being more widely used. Although the quantity of material is the same for all pupils examined, and the difficulty of the work increases by fairly even steps, both the time and the quality of the reading are allowed to vary at the same time, which makes the results less convincing than they otherwise might be.

Thorndike's Visual Vocabulary Scales. — For practical purposes one may consider these scales as four forms of the same scale, although the publishers have listed them as Scale *A2x*, Scale *A2y*, Scale *Bx*, and Scale *By*.[1] The purpose of these tests is to determine the extent of the vocabulary possessed by the class or grade group. These scales were devised for measuring pupils in grades 4 to 8, but they are useful in the third elementary grade and in the first

[1] For a more complete discussion of these scales see the writer's booklet entitled, "Directions for Using the Thorndike Visual Vocabulary Scales," Teachers College, Columbia University.

two high-school grades as well. Each scale contains from 10 to 13 lines of words to be classified, each line containing ten words of the same difficulty. The score in this test is in terms of the difficulty of the words successfully classified. The general directions and four lines (having difficulty values of 4, 6, 8, and 10, respectively) from Scale *A2y* are reproduced here.

SELECTED SAMPLES FROM THORNDIKE'S VISUAL VOCABULARY SCALES

Write the letter F under every word that means a *flower*.

Write the letter A under every word that means an *animal*.

Write the letter N under every word that means a *boy's name*.

Write the letter G under every word that means a *game*.

Write the letter B under every word that means a *book*.

Write the letter T under every word like *now* or *then* that means something to do with *time*.

Write the word GOOD under every word that means something *good to be* or *do*.

Write the word BAD under every word that means something *bad to be* or *do*.

4y. wolf, lily, bear, kind, clean, buttercup, cruel, truthful, elephant, baseball

6y. during, noon, jesse, month, jeffrey, oscar, forenoon, generous, later, rhinoceros

8y. croquet, jasper, capable, hollyhock, jason, shuffleboard, conrad, owen, gradual, hiram

10y. anemone, depraved, beagle, balsam, hugh, prudent, calliopsis, whist, tolerant, roderic

From a statistical point of view this is one of the best of the reading tests. It measures the difficulty of the words that one can in a given time identify well enough to write the correct notation. The amount done by each pupil also varies, it is true; but since there is an equal number of words at each level of difficulty, the number identified is a function of their difficulty. The time limit is so long that pupils have more than sufficient time to identify all the words they know. The dominant and important variable is difficulty, which makes this an excellent test (statistically) of the pupil's power to read or classify words.

There are certain minor criticisms of this test which have kept it from being used as extensively as it deserves. The method of calculating scores is somewhat complicated, although the resulting scores are more reliable and meaningful than the scores in the average test. Certain words like " samuel " and " daniel " may be classified in two or more ways, either one of which would be right. Many critics have also claimed that when begun with a small letter the above are not words at all and cannot be classified as proper names. It would perhaps have been better had Thorndike printed every word in the list entirely in capitals, thus eliminating the last-mentioned criticism. His recent Test of Word Knowledge will probably be more popular than these earlier vocabulary scales have been.

The Thorndike-McCall Reading Scales. — There are ten forms of this series of scales, which are intended to measure the degree of reading difficulty in sentences and paragraphs that pupils can master perfectly in a given time. Although there are not exactly the same number of paragraphs and questions in each scale, and the difficulty

of a question is not known to be exactly the same as the difficulty of the corresponding question on other forms, Professor McCall has provided with each form a table for converting " raw scores " into scores on a " T Scale," which is statistically the same for all forms. Theoretically a pupil's score on the T Scale would be the same regardless of the form of the test on which he is measured. The units of the T Scale are based upon the difficulty of the work done, which causes high scores to indicate that great difficulty was successfully met and low scores to signify that less difficulty was overcome. Samples from Form 4 of this scale are given below.

SELECTED READINGS FROM THE THORNDIKE-McCALL READING SCALES

Read this and then write the answers. Read it again if you need to.

Fred lives in the country. He likes to hunt and fish, and has a gun that cost sixteen dollars. His sister Grace keeps hens and ducks, and sells the eggs. She is learning to play the piano, and goes to Miss Thomas for a lesson every Saturday. She likes music but Fred doesn't.

1. Write the name of the boy who is fond of hunting . .
   . . . . . .
2. Write the name of the child who likes playing the piano
   . . . . . .
3. Who is learning to play the piano? . . . . . .
4. Where does the boy who has the gun live? . . . . . .
   . . . . . .

Read this and then write the answers. Read it again if you need to.

Inside the big box made of wood and iron that stands under the bed there is a box just like it, only smaller, and a brass key to open the small box. Whoever opens it will find three keys, one of gold, one of silver and one of steel. The first opens the door of the red room, the last opens the door of the blue room. The other opens the outside gate. Every key has a ring.

16. Where will you find the big box? .....

17. What is the chief difference between the boxes? ....

.....

18. What will you find on each of the rings? .....

This test is used in all grades from the third up through the high school. Since answering more questions means doing more difficult reading (amount done having been made a function of increasing difficulty, as in the vocabulary scales), these scales are excellent for the measurement of general reading power. It must be admitted, of course, that the results from these tests do not point out what particular phases of ability in reading need to be emphasized by the teacher, any more than the result of measuring a child's general weight tells what to feed him. The Thorndike-McCall Test is furthermore " so constructed that the thing measured is the ability of the pupil to use reading for a specific purpose." It may be worth while to measure the type of reading which pays " little attention to precision of understanding " and is characterized " by a passive attitude toward what is read," but this test undertakes

to hold the quality of the reading constant, — at a standard which is sufficiently good to enable one to make use of what one reads.

One danger in the use of this scale arises from the very complete booklet of directions accompanying it. With the aid of this booklet one converts the actual number of questions answered correctly into a T Score, which may then be converted by means of another table into a Reading Age, which may then be converted into a Reading Quotient, which (theoretically) may then be compared with an Intelligence Quotient in order to obtain an Efficiency or Accomplishment Quotient. In other words, the directions and tables tend to make one forget the crudity of the original scores and to put more confidence in the exactness of the derived scores than they deserve. If a child happened to answer 23 questions correctly in Form 4, for example, he is to be counted as having a Reading Age of 12 years, 0 months, while if he had answered one more question correctly his R. A. would have been 12 years, 8 months. This type of danger is especially likely to overtake the thoughtless or the inexperienced examiner.

The Burgess Scales for Silent Reading. — In many respects the scales prepared by Mrs. Burgess may be considered the best yet available for measuring certain aspects of reading ability.[1] Each scale consists of twenty equally difficult paragraphs, each of which is to be read well enough to enable the reader to do what it tells him to do. The variable measured is the number of these paragraphs read and obeyed in five minutes. This series of scales bears the

[1] M. A. Burgess, *The Measurement of Silent Reading*, Russell Sage Foundation.

same relation to reading ability as the Courtis Tests bear to arithmetic ability. They measure the rate at which one can do equally difficult tasks. When one has obtained a general view of the power possessed by a class in reading, through the results from a scale measuring difficulty, he may then begin an analysis of the situation by the use of such rate tests as the Burgess Scales and such vocabulary tests as Thorndike's.

Any criticism of these scales would probably be equally applicable to all rate tests. It requires only a little practice in rapid reading to enable the child to make a decided improvement in his score on a rate test. Instances have come to the attention of the writer in which pupils have doubled their scores through no other practice than taking four or five forms of the Burgess Scale. The scale will be very useful, however, in calling attention to those pupils who are unable to make large gains in speed because of defective eye-movements and other serious defects. Some teachers have objected to the scale on the ground that a careful child is penalized for following directions closely, — trying, for example, to make artistic footprints in the snow and pretty feathers on the ground. This can be avoided by instructing the pupils before they begin work not to try to draw beautiful pictures.

The Gray Silent Reading Tests. — This test attempts to obtain a larger amount of information about the pupil's reading than the others mentioned in this volume, but in order to secure this information only one pupil can be tested at a time. Three selections are used as material for this test, so it may for practical purposes be considered as having three different scales: one for the 2d and 3d grades;

another for the 4th, 5th, and 6th grades; and a third scale for the 7th and 8th grades.

The most unique feature of this test is the ingenious method by which the examiner is able to obtain a measure of the rate at which the child naturally reads a story. Each selection to be read is divided into three parts, which are printed side by side on a wide cardboard. The child is told to read the story from beginning to end without stopping. The examiner watches the pupil closely and starts his stop-watch when the child raises his eyes to begin the second section or page of the story. He stops his watch or records the elapsed time when the pupil looks up to begin the third section. This gives an accurate measurement of the time normally required by the pupil to read a given amount, without making the pupil conscious of the fact that he is being timed. Many pupils are very sensitive to having their work timed, and this device is the best the writer has seen for obtaining natural time records.

The directions to the pupil in this test are as follows:

"Read the story on this card silently. Read the story from beginning to end without stopping or repeating any of it. Read the story rapidly but carefully. Do not stop reading to ask about difficult words; read such as best you can. Be prepared to tell the story or to answer any question about it when you are through. Do you understand?"

For 2d- or 3d-grade pupils the next directions, after the story has been read, are as follows:

"Tell me the story of Tiny Tad as well as you can. Tell just those things which you read in the story. Tell it to me slowly so I can copy it. Do you understand?"

In the 4th grade and above, the pupils write their own reproduction of the story, and then write their own answers to the ten printed questions. In the lower grades, however, the examiner writes the story as the child tells it and then copies the child's answers to the questions.

The reproduction of the story given by the child is scored according to the percentage which the number of words in the child's story is of the number in the original selection (after repetitions, wrong and irrelevant statements have been removed from the child's account). Each answer the child gives correctly is to be rated ten points. The score on the ten questions is then to be added to the percentage for the child's reproduction of the story, and the total divided by two, which will give a number or value to be called a Quality Score. The Time Score for each pupil is the number of seconds required to read 100 words.

It is clear that this test violates in some respects the principle of a dominant variable, for both time and quality of reading are allowed to vary without the one being subordinated and proportional to the other. It is also doubtful whether the questions and the reproduction deserve equal weight in determining the quality of the reading, and whether each question is of exactly the same importance as each other question. There would also have been a distinct advantage in using exactly the same reading materials for all grades in order that progress might have been more accurately measured. Many teachers object to the test, however, on the less important ground that it requires too much time. The test deserves wide use, if only for the purpose of determining the rate at which each pupil normally reads an interesting story.

The Courtis Silent Reading Test No. 2. — With the Courtis Silent Reading Test No. 2, an entire class of pupils is tested at the same time. This test is in two parts, but both parts are to be given to all pupils from grades II to VI inclusive.[1] In the first part, the pupils read a two-page story for three minutes, making every thirty seconds, at a signal from the examiner, a mark around the last word read. In the second part of this test the children read again the same story, but this time in fourteen short sections, and then write " yes " or " no " in answer to five questions about each section. In the second part the examiner calls " Mark " every sixty seconds and stops the work at the end of five minutes.

The device by which the pupils make marks to indicate the rate at which they read proves distracting to some pupils, especially in the lower grades, although most of them soon gain the ability to make these marks without being disturbed by the examiner's command to " Mark." Part II of the test assumes without proof that all of the 70 questions are of the same difficulty, and attempts to obtain an index of comprehension by relating the number of correct answers to the total number of answers. Part I gives a measure of the rate at which the pupils read and mark, without any check on how well they read, and with the assumption that one part of the story is of the same difficulty as any other part. Mr. Courtis states that this test measures " simply the ability to read silently and understand a simple story and simple questions about the story." There are three forms of the test, although each should

[1] The writer would not advise the use of this test below the fourth grade.

probably be considered as an individual and distinct scale until it has been proved that they are equivalent.

Other reading tests. — There are many other reading tests available, several of them being superior in certain respects to some of those described. The writer has chosen for presentation here only a few that represent certain types of test procedure.

The Haggerty Reading Examinations are interesting to the pupils and quite comprehensive, but they are so complex and measure a mixture of so many different phases of reading ability that they may almost as well be called measures of intelligence as measures of reading ability. The writer approves of this variety and comprehensiveness in a reading examination, when it is recognized as a test of general reading ability, and when no attempt is made to draw elaborate inferences and conclusions about special phases of reading ability from the scores obtained.

The Kansas Silent Reading Tests [1] have probably been as widely used as any other reading tests, and yet they are extremely defective in their construction and in their interpretation. The values of the different elements in the tests were determined by the relative amounts of time required to perform the tasks rather than by the relative number of persons able to perform them. The amounts of difference between consecutive elements vary so greatly that a false impression of the amounts of difference between the abilities of pupils is created. This erroneous impression with regard to the differences between pupils is further aggravated by a cumulative method of obtaining a pupil's

[1] Bulletin No. 3, Bureau of Educational Measurements and Standards, State Normal School, Emporia, Kansas.

score.[1] The Monroe Standardized Silent Reading Test employs the same defective method for obtaining final scores.

The Briggs English Form Tests should probably be mentioned here as excellent instruments for measuring what practically amounts to a special type of reading ability — proof reading.[2] These are particularly valuable in the junior and senior high-school grades. The test not only measures difficulty of performance as its dominant variable, but it is so arranged as to give specific information about the special types of errors made by the pupils. It would be useful if similar tests were available for the pupils of the lower grades.

Summary. — One of the questions most frequently asked of every worker in educational measurements, is " What is the best test in reading? " The only correct answer to this question is that there is no " best reading test." Any one who asks for a " best reading test," or for a " best arith-

[1] In Test I (for grades 3, 4, and 5), for example, the first two elements have a value (according to the author), of 1.2, the twelfth has a value of 2.2, and the thirteenth has a value of 3.1 points. The difference in value between doing one task and doing the first twelve tasks is therefore 1.0, while the difference between doing twelve and doing thirteen is .9 of one unit. By the cumulative scoring used in this test, the difference between doing one and doing the first twelve is 19.5, while the difference between doing twelve and doing thirteen is 3.1 points. Practically the same amount of improvement secures only one-sixth the amount of gain in score at one place on the scale as at another place.

[2] See "An English Form Test," by Dr. T. H. Briggs in the *Teachers College Record*, Vol. XXII, No. 1. This article may be obtained separately from the Bureau of Publications, Teachers College, Columbia University.

metic test" thereby betrays his limited knowledge of the purposes and uses of educational measurements. Measurements are a means for answering definite questions exactly. It is desirable in every case to choose the scale that will best measure that ability or phase of ability about which one wishes to know. If one is interested in knowing how rapidly pupils can do equally difficult reading tasks, he should use the Burgess Scale, or something like it. If one is interested in knowing how difficult a task in reading the pupils can do, he should use the Thorndike-McCall Scale, or something like it. The time has arrived in the history of educational measurements when supervisors and teachers should choose with intelligence the tests they use in the schoolroom. There has been entirely too much giving of this or of that test merely to " be in style."

QUESTIONS AND EXERCISES

1. How does the U. S. Census define literacy? Does this standard mean the same to all census-takers?

2. For what reasons is it a serious matter that more than a fifth of our population is practically unable to read the English language?

3. Certain Sunday Schools give prizes to the boys and girls who learn the largest number of Bible verses. Are the units of amount equal in such a contest? The units of difficulty? The units of time? The units of quality or excellence?

4. If a superintendent of schools wished to discover how his school compared with others in general reading ability, would it be better to use the Gray Oral Reading Test or the Gray Silent Reading Test? Why?

5. If a principal wished to compare two fifth-grade groups as to general reading ability, would it be better to use a Thorndike-McCall Reading Scale or a Thorndike Visual Vocabulary Scale? Why?

6. If a teacher wished to select the pupils from her class who were most in need of special drill in rapid reading, what test should she use?

7. If a teacher had given the Haggerty Reading examination and discovered from the results that ten of her pupils were much behind the other members of the class in general reading ability, what further tests might be given to these ten in order to discover the specific difficulties that needed correction in each pupil's reading?

8. What elements of reading ability are at the present time least adequately measured? How might tests be arranged in order to measure each of these elements?

REFERENCES FOR READING

M. A. Burgess, *The Measurement of Silent Reading*, Russell Sage Foundation.

C. T. Gray, *Studies of Elementary School Reading through Standardized Tests*, University of Chicago Press.

Judd and Buswell, *Silent Reading: A Study of the Various Types*, University of Chicago Press.

W. A. Smith, *The Reading Process*.

C. R. Stone, *Silent and Oral Reading*.

CHAPTER XVI

SCALES FOR QUALITY IN ENGLISH COMPOSITION

HONESTY is a virtue which one may assume to be present in teachers of English composition as frequently as in teachers of any other school subject. Yet there are probably greater differences between the marks assigned to the same composition by different English teachers than between the grades assigned to a paper in any other school subject by the different teachers of that subject. Apart from obvious mechanical errors, the only way to prove that one selection of art, music, or literature is better or worse than another selection is to have the two selections compared by a number of competent judges. In other school subjects the mechanical elements are usually so much greater than the artistic that a reasonable amount of agreement may naturally be expected, but in English composition the artistic elements predominate, at least in the higher grades. Only in so far as they have compared notes and adopted common objective standards should one expect the judgments of English teachers regarding the quality of a composition to approach agreement.

As suggested in Chapter XIV, it might be possible to measure the quality of English compositions by means of a score-card on which all of the characteristics of a perfect

composition had been listed in detail and assigned a definite proportion of the total number of points. In employing such a score-card, one would examine a child's composition for a number of specific details, assigning to each detail a certain fraction of the maximum number of points allowed for perfection in that item; then one would sum up all of the points allowed on the individual items, and obtain a final score for the total composition. The usefulness of such an analysis of perfect composition has not been questioned, but few teachers are competent to examine each composition for a score or more of small details, and still fewer are willing to give the time required for such a task. The only scheme which has attained any wide popularity is to have a graduated set of samples with which to compare the general quality of the pupil's composition. In using this type of scale one finds a sample in the standard set which seems to have approximately the same general quality as has the pupil's composition. The pupil's composition is then assigned the same numerical or letter value as its equal in the scale.

The composition scale, instead of describing a perfect English composition, reproduces one very nearly perfect, another slightly less perfect, and still others that are less and less perfect, down to a composition which seems not to have any merit at all. Instead of asking the teacher or judge to estimate how much less perfect certain details are than a described perfection, the scale requires one to compare the general quality of the pupil's work with the general quality of the compositions in the standard series. One merely says, "In general it seems about equal to number 4." This is much easier than having to say,

" In punctuation it seems about half perfect, in capitalization about nine-tenths perfect, in paragraphing about one-tenth perfect, in interest about three-tenths perfect, etc."

The Hillegas Composition Scale. — The first carefully prepared scale for measuring the quality of English compositions was published by Dr. Milo B. Hillegas in September, 1912.[1] This scale contained ten compositions, each on a different subject, and each decidedly better than the one just before it. These ten compositions were printed on a long narrow strip of paper, the poorest one at the top and the best at the bottom. Although the increases in the amounts of quality were not exactly the same as a reader passed from one sample to the next better one, they were as nearly the same as it was possible to get from the materials used. The derivation of scales will be discussed in detail in Chapter XX, but it is worth while to state here that the unit employed for measuring quality was " such a difference as could be noticed by exactly three-fourths of the 202 judges." Four of the ten samples in the Hillegas Scale are reproduced below:

Sample 580. **Value 0.** Artificial sample.

Letter.

Dear Sir: I write to say that it aint a square deal Schools is I say they is I went to a school. red and gree green and brown aint it hito bit I say he don't know his business not today nor yeaterday and you know it and I want Jennie to get me out.

[1] M. B. Hillegas, "A Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record*, Vol. 13, No. 4.

Sample 94. **Value 369.** Written by a boy in the second year of the high school, aged 14 years.

Sulla as a Tyrant.

When Sulla came back from his conquest Marius had put himself consul so sulla with the army he had with him in his conquest siezed the government from Marius and put himself in consul and had a list of his enemys printy and the men whoes names were on this list we beheaded.

Sample 196. **Value 675.** Written by a girl in the first year of the high school, aged 18 years.

Ichabod Crane.

Ichabod Crane was a schoolmaster in a place called Sleepy Hollow. He was tall and slim with broad shoulders, long arms that dangled far below his coat sleeves. His feet looked as if they might easily have been used for shovels. His nose was long and his entire frame was most loosely hung to-gether.

Sample 571. **Value 838.** Written by a boy in the Freshman class in college.

Venus of Melos.

In looking at this statue we think, not of wisdom, or power, or force, but just of beauty. She stands resting the weight of her body on one foot, and advancing the other (left) with knee bent. The posture causes the figure to sway slightly to one side, describing a fine curved line. The lower limbs are draped but the upper part of the body is uncovered. (The unfortunate loss of the statue's arms prevents a positive knowledge of its original attitude.) The eyes are partly closed, having something of a dreamy langour. The nose is perfectly cut, the mouth and chin are moulded in adorable curves. Yet to say that every feature is of faultless perfection is but cold praise. No analysis can convey the sense of her peerless beauty.

It will be obvious to the thoughtful reader that such a scale is wholly independent of any particular test procedure. It would be possible to measure with this scale the quality of a composition written in ten minutes as easily as to measure the quality of a composition written in sixty minutes. This scale measures the general quality of the result, regardless of the difficulty or the time of the process, and regardless of the amount written. In order to make just comparisons between the work of pupils in different places, test conditions must be set up to control these other variables. Ability in English composition could not be measured merely by determining the quality of the composition some pupil had written, for this pupil might have had an easier subject and a much longer time to work than pupils elsewhere had. Hudelson has discovered, for example, that the average pupil writes compositions more than one full unit poorer when writing on the subject, "What I Should Like to Do Next Saturday," than he writes when the subject is, "How I Learned a Lesson." The Hillegas Scale is not a test. It is only a scale for measuring the quality of compositions, which may or may not have been produced under standard test conditions.

Several limitations in the Hillegas Scale became apparent soon after it was published. In fact, the author of the scale pointed out most of them at the time of publication. Several of the first compositions in the scale were written by adults who were trying to write "very poor English" in order to supply samples that were poor enough to meet the statistical requirements. Several of the samples were so short as to give one little opportunity to judge their real worth. The wide differences in the subject matter of the

various samples in the scale made it difficult to distinguish wherein each was superior to the previous sample. There were also users of the scale who felt that one sample of a given quality was not enough to supply a clear impression of just how excellent it was. The amounts by which each example differed in quality from the previous example were unequal in different parts of the scale. The essential validity of the scale was nevertheless recognized by those who gave it serious thought, and many modifications of it have since been produced.

Thorndike's Extension of the Hillegas Scale. — The supplement or extension of the Hillegas Scale prepared by Thorndike is an improvement chiefly in that it offers a larger number of samples at several points on the scale. The parallel columns below show how the values of the samples are distributed in the two scales.

Hillegas Scale Values	Thorndike Extension Values
0	0
1.8	1.8
2.6	
	3.0
3.7	3.7
	4.0
4.7	4.7
	5.0 5.0
5.8	
	6.0 6.0 6.0 6.0
6.7	
	7.0 7.0 7.0 7.0 7.0
7.7	7.7
	8.0 8.0 8.0 8.0 8.0 8.0
8.4	8.4
	9.0 9.0
9.4	9.4
	9.5

This extension also has the advantage of containing several compositions on the same subject, which makes it somewhat easier for one to see the differences between samples having different quality values. Any one who is planning to make extensive use of the Hillegas Scale in any form should spend a considerable amount of time studying and working with this supplement, for the sake of the familiarity it would give with the significance of qualities 5, 6, 7, 8, and 9.

The Nassau County Extension of the Hillegas Scale. — In making an inventory of the schools of Nassau County, New York, it became evident that children actually do write compositions of very low quality. This fact and the difficulty encountered by most teachers in judging the values of the artificial samples at the low end of the Hillegas Scale, led the writer to undertake to supplement the scale by a new series of samples, all of which should be of the same general narrative character. The Nassau County Supplement is therefore a second series of ten samples of English composition. Every alternate composition in this supplement is reproduced below.

ALTERNATE SAMPLES FROM NASSAU COUNTY SUPPLEMENT TO THE HILLEGAS SCALE

0. .004	What I should like to do next Saturday I went going on to the Dox Saturdaye dnd day we the boys and I well going home and I well going the boys. and I will going these read in and they to night. and we or night. I well going a ground shalt and I gone out I will going to shea shouse and I will shoe or the skill of the shea of night.

1.9

1.93

one next S aturday I expect to go to the city leve next G aturday to see my ofriend archie king I am going to grow to the baning balys circus with hime next S aturday fefore I go I have to do my jobs feedsing the cows ard horse ard chinkens and geese next Saturday
My friend is a very good fellow to go and see So my mother S aid " If I do my work during Easter week vacation I can go to the barning baley circus with. hime

3.8

3.84

I would like to go out in the after noon and play catching the ball. Go over to Bertha's house and have a few girls to come with me and be on each others side. I have a tennis ball too play with. The game is that one person should stand quite aways from another person and throw the ball too one then another. Someone has to be in the middle and try too get the ball a way from someone then she takes this persons place who she caught the ball from. Then till every person has a chance.

6.0

6.01

I should like to join my girl friends, who are going to the city on the 9 :05 A. M. train. They are going shopping in the morning and will have lunch to-gether, then they are going to the Hippodrome. After the Hippodrome, they are all going home to dinner to one of the girls houses, she lives on Riverside Drive so they expect to take the " Fifth Avenue Bus " up there. The evening will be devoted to playing games, singing and dancing.

8.0

One Sunday, towards the end of my summer vacation, I was in bathing at the Parkway Baths. In the Brighton Beach Motordrome, a few rods away, an aviation meet was going on. Several times one of the droning machines had gone whirring by over our heads, so that when the buzzing exhaust of a flier was heard it did not cause very much comment. Soon, however, the white planes of " Tom " Sopwith's Wright machine were seen glimmering above the grandstand. Everyone stood spellbound as he circled the track several times and then headed out to sea. He was seen to have a passenger with him. Suddenly, the regular hum of his motor was broken by severe pops, and the engine ran slower, missing fire badly. In response, to Sopwith's movements, the big flier tilted and swooped down to the beach from aloft like an eagle. The terrified crowd made a rush to get out of the way as the airship came on, but Sopwith could not land on the beach, but skimmed along close to the water instead. Suddenly his wing caught the water, and the big machine somersaulted and sank beneath the waves. The aviators soon came bobbing up and were taken away in a launch, but the accident will not soon be forgotten by those who saw it.

8.00

In addition to the statistical work necessary to establish the new scale, the writer suggested a uniform test procedure to accompany the scale and set up tentative standards of achievement on this test.[1] These standards and

[1] M. R. Trabue, "Supplementing the Hillegas Scale," *Teachers College Record*, Vol. 18, No. 1.

the values of the samples in the supplement were expressed in terms of Hillegas Scale values and were determined experimentally. The average results that have actually been obtained by the use of the test are not as high as the proposed standards.

The Virginia Supplement to the Hillegas Scale. — This supplement, which is now published as the Hudelson English Composition Scale, was prepared by Professor Earl Hudelson from compositions written by public school pupils for the school survey in Virginia. The topic on which the majority of these compositions were written is, "The Most Exciting Ride I Ever Had." The values in this scale are in terms of the original Hillegas Scale. One may therefore compare the scores obtained by its use with those obtained from the Hillegas Scale, the Nassau County Supplement, or the Thorndike Extension of the Hillegas Scale. The Virginia Supplement differs from the others in having a remarkably even spacing between the qualities of consecutive samples. The sixteen compositions included in this supplement have approximately the following values: 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, 9.0, and 9.5.

Professor Hudelson has recently conducted some very careful investigations of ability in English composition and on the basis of his findings has constructed both a test and a scale for measuring the typical or average achievements of pupils in this field.[1] The Typical Composition Ability Scale has one sample at 0, another at 1.0, a third at 2.0, and so on, up to 9.0, with extra samples at 3.0, 4.0, 5.0, 6.0, and

[1] Earl Hudelson, *English Composition: Its Aims, Methods and Measurement*, pp. 80–99.

7.0, which are the values most frequently used. This scale is one of the most useful derivatives of the Hillegas Scale.

Which form of the Hillegas Scale to use. — If the ratings were made with equal accuracy on all forms, a given class of pupils would probably obtain the same average score and the same measure of dispersion, whether their compositions were rated on the original Hillegas Scale or on one of its supplements. Experience has demonstrated, however, that it is not equally easy to be accurate in the use of the various forms of the scale. The suggestions made here are based upon wide experience in the use of the scales and in teaching others to use them, but they are merely suggestions and should not be considered as absolute rules.

The unit of difference used in these scales is "such a difference as three out of four persons would see." The fourth judge would not see the difference between compositions whose qualities were one unit or less apart, unless he happened to have had training in the rating of compositions. In a group of untrained teachers it would probably be found that, if not told their values beforehand, 25% of them would say that sample 7.7 was poorer than sample 6.7 on the Hillegas Scale. An equal proportion of them would probably say that sample 6.0 of the Nassau County Supplement was poorer than sample 5.0. A difference which 25% of the teachers cannot see is about as small a difference as one would ordinarily care to use.

Actual experience has demonstrated that it is desirable to begin with even larger differences than one unit of quality. The writer has had best results from beginning groups of teachers by mimeographing for them every alternate sample on the Nassau County Supplement, which gives

them the same scale as appears on page 369. The differences between consecutive samples here are two units of quality. Such a difference in quality can be seen by at least nine out of ten untrained persons. A similar and equally useful beginner's scale could be prepared by taking qualities 2.0, 4.0, 6.0, and 8.0 from one of Hudelson's scales.

On such a short scale, beginners can be quickly taught to rate compositions to the nearest sample. After a little practice with this scale, they can be given the entire Nassau County Supplement, or some other one-unit-step scale, and again taught to rate pupils' compositions to the nearest scale sample, just as with the shorter scale. After extensive practice and checking of their judgments, they may take the Hudelson Scale as the basis for finer work. The half-unit steps of the Hudelson Scale are so short that 37% of a group of untrained persons were unable to recognize the differences between consecutive samples. Extended training and experience make it possible for one to use the Hudelson Scale to advantage, however, and it is recommended especially for accurate scientific work with well-trained judges.

The Thorndike Extension, with its wealth of material at frequently used points, is exceedingly useful for the training of judges who are to do accurate work in some scientific experiment. It answers admirably such questions as "Just how good is quality 6.0, quality 7.0, or quality 8.0?" It is recommended that any one who is planning to use some other form of the scale for extensive work should first make a careful study of Thorndike's Extension. This extension is often useful as a reference against which to check doubtful decisions.

How to rate compositions on the Hillegas Scale. — Whatever form of the Hillegas Scale is being used, it is statistically desirable that all ratings should be to the nearest sample on the scale. It is possible for well-trained observers to decide that a given composition is worth more than 6.0 and less than 7.0, or even to feel that its true value is probably 6.7, but it is not usually desirable to make such interpolations. When one becomes such an expert that he feels confident of his ability to interpolate accurately between the values appearing on the Hillegas Scale or on the Nassau County Supplement, he should change scales and use the Hudelson Scale. It is better to use a finer scale and continue rating to the nearest sample than to assign values intermediate between those on the scale.

One should also be careful to make his comparisons equitably. It is best to pass up the scale first, from the low end, comparing the quality of the child's composition with the quality of those on the scale, until a printed sample is finally found which is distinctly better than the one the child has written. Then one should begin above this point and come down the scale until a sample is found which is distinctly poorer than the child's composition. A compromise may then be worked out between these two values, giving to the pupil's work the numerical value of the scale-sample to which it is most nearly equal in general merit.

No attempt should be made to reduce the matter of general composition quality, illustrated in the various forms of the Hillegas Scale, to a mathematical basis of so many omitted commas, so many short sentences, so many misspelled words, so many well-used adjectives, and the like. Such an attempt would not only be difficult to ac-

complish successfully, but it would require too much time to rate compositions on this basis. From the point of view of measurements, more useful results will be obtained by having each production rated by two or three different judges for its general value as an English composition, rather than by counting errors and excellences. Second or third ratings of the same theme should always be made without knowledge of the previous ratings.

The reader should not from the above discussion decide to judge the composition work of pupils by general impressions only. What he should consider is general merit. For the purposes of instruction the teacher may wish frequently to grade the child's daily work on the basis of such details as capital letters or commas. Measurement and instruction are not the same process, however. A test for general quality in English composition does not improve the quality it measures any more than weighing a child makes him heavier.

A standard test in English composition. — Compositions may be rated on the Hillegas Scale in order to measure the effect of different topics, different time limits, different examiners, or for any other purpose for which measurements might be desirable. If one is to compare schools or children as to their efficiency in composition, however, the same time limits, the same topics, and the same general directions are necessary. In other words, efficiency in composition cannot be measured by the quality of the product only. Quality may be the dominant variable in such measurement, but standard test conditions must be provided to control the other three variables. The directions for securing samples of handwriting, printed on pages

173 and 174, will be found quite adequate for this purpose, although Hudelson's Typical Composition Ability Assignment is perhaps better. At least three persons should rate each composition, each person making his rating without any knowledge of the marks assigned by the others. The median or middle rating of the three judges may be taken as the one to represent the final value of the composition. In case the first two marks are identical, the third judgment is not so necessary.

Standard scores in English composition. — The composition test is one of those useful tests which may be given to pupils at any stage of advancement. The Hillegas Scale extends down far enough to measure the poorest composition any child may write, and up far enough at the other end to measure almost any composition that may be written in the high school. This enables one to secure an extensive classification of scores by ages within each grade, which seems to the writer so generally desirable. Table LI makes a comparison of the results obtained in St. Paul, Minnesota, and in Baltimore, Maryland, as presented by the school survey reports for these two cities.

The information supplied by Table LI is more complete than that supplied by mere grade averages. The tabulations used in preparing it supply also without extra labor the data for both age and grade scores as in Table LII. It is unfortunate that many of our educational tests cover only a limited range of ability and can be used in only a few school grades, making such an analysis as Table LI impossible. The tendency, clearly observable in Table LI, for older pupils in a grade to lower the result when only grade averages are to be compared, would seem to penalize

TABLE LI

Comparison of English Composition Results by Age within each Grade

Grade and Age	St. Paul	Baltimore
Fourth Grade		
Age 8	—	2.5
9	2.0[1]	2.7
10	1.9	2.6
11	2.1	2.6
12	2.0	2.6
13	1.9[2]	2.5
14	—	2.3
15	—	2.2
All ages	2.0	2.6
Fifth Grade		
Age 9	—	3.8
10	3.4[1]	3.5
11	3.4	3.4
12	3.2	3.3
13	3.4	3.2
14	3.3[2]	3.1
15	—	2.9
All ages	3.4	3.3
Sixth Grade		
Age 10	3.2	4.1
11	3.8	4.2
12	3.5	3.9
13	3.3	3.8
14	3.5	3.6
15	3.3	3.2
All ages	3.5	3.9
Seventh Grade		
Age 11	4.3	5.0
12	4.2	4.7
13	4.1	4.8
14	4.0	4.8
15	4.0	4.4[2]
16	3.9	—
All ages	4.1	4.8
Eighth Grade		
Age 12	5.0	5.6
13	5.0	5.7
14	4.9	4.5
15	4.8	4.8
16	4.7	5.5
All ages	5.0	5.4

[1] Includes those of the age indicated or younger.
[2] Includes those of the age indicated or older.

TABLE LII

COMPARISON OF ENGLISH COMPOSITION RESULTS BY AGE AND BY GRADE
MEDIAN SCORES ON THE HILLEGAS SCALE

AGE OR GRADE	MEDIAN SCORES OF ELEMENTARY SCHOOL PUPILS		
	In St. Paul	In Baltimore	In Nassau County
Grade IV	2.02	2.59	2.76
V	3.38	3.34	3.42
VI	3.54	3.91	3.82
VII	4.12	4.77	4.18
VIII	4.96	5.43	4.56
Age 9	2.04	2.75	2.91
10	2.83	3.05	3.10
11	3.50	3.45	3.40
12	3.95	3.64	3.70
13	4.50	4.01	3.97
14	5.02	4.30	4.42
15	5.54	4.18	—
16	5.96	5.55	—

the class or school which retained its over-age pupils in school. A teacher may raise the median composition scores for her class by eliminating the older pupils from school, as well as by teaching them how to write better compositions.

Two sets of standard scores in English composition are given in Table LIII. The first is based on estimates by English teachers and supervisors of what might naturally be expected of pupils in each grade, and the second is based on what pupils in each grade actually accomplish.[1]

[1] Those who wish to compare with specific school systems will find a complete report in the *Twenty-Second Yearbook of the National*

TABLE LIII

STANDARD ENGLISH COMPOSITION SCORES BY GRADES
(HILLEGAS SCALE)

SCHOOL GRADE	IV	V	VI	VII	VIII	IX	X	XI	XII
Desirable average	3.5	4.0	4.5	5.0	5.5	6.0	6.5	6.9	7.2
Average, all ages	3.0	3.6	4.2	4.7	5.3	5.5	5.9	6.3	6.7

The scores in Table LIII are for pupils in the average American public school, at the middle of the school year. Some allowance would need to be made in each grade for a school which was composed largely of foreign-speaking children. A technical or manual arts school should perhaps be allowed a slightly lower standard. If pupils are examined at the beginning of the year rather than at the middle of the grade, still further allowances should be made.

The Harvard-Newton Composition Scales. — Shortly after the appearance of the Hillegas Scale, Dr. Frank W. Ballou, who was then in charge of educational measurements in the Boston Public Schools, undertook to prepare special scales for measuring the quality of compositions written by eighth-grade pupils. Instead of having one scale for measuring the general merit of all types of compositions, Dr. Ballou proposed to have one scale for Description, one for Narration, one for Exposition, and one for Argumentation. The scales finally prepared were based on the judgments of about twenty-five teachers in

Society for the Study of Education, Part I. The writer is indebted to Professor Hudelson for permission to use his composite standards, — the second line of scores in Table LIII.

the Newton Public Schools and were published by Harvard University.

Each of the Harvard-Newton scales consists of six compositions, rated A, B, C, D, E, and F, respectively. Each of the six compositions is accompanied by a statement of its peculiar *Merits*, another statement of its *Defects*, and then by a statement of how it compares with the one just better and with the one just poorer. This discussion of each composition undoubtedly is an aid to teachers who have difficulty in discovering wherein the differences lie.

Although most of the textbooks on English composition note and emphasize three or more types of compositions, it is frequently difficult to decide which of the Harvard-Newton Scales should be used on a child's product. Furthermore, a grade of C does not mean exactly the same for one of these scales as for another, so that one must specify in each record the particular scale used. The use of twenty-five judges in establishing the scales gives them more weight than using a single judgment would have given, but leaves them with less reliability than the Hillegas Scale, which was standardized on more than two hundred judges' ratings. The greatest defect in these scales is the fact that they do not extend down far enough to measure the poorest products of even the ordinary eighth grade. Although built for measuring eighth-grade work, entirely too large a proportion of the compositions written by pupils in this grade (or in lower grades) must be scored F.

The Minnesota English Composition Scales. — Dr. M. J. Van Wagenen of the University of Minnesota has prepared three scales for measuring the quality of English compositions. One scale is a Narration Scale, one is a

Description Scale, and one is an Exposition Scale. Each scale contains about fifteen compositions, each of which has a different mark for each of three qualities: Structure, Mechanics, and Thought Content. The value for structure is not the same as the value for Mechanics, nor as the value for Thought Content. In using this scale one would decide first of all whether a pupil's composition was narration, description, or exposition. It would then have to be rated on the appropriate scale, first for its sentence and paragraph structure, then for its mechanical errors, and finally for its thought content. It is not probable that this series of scales will be used widely outside the State of Minnesota, because of the complexity of the task of making ratings on it.

The Willing Scale for measuring written composition. — This scale may be termed "a one-man scale," since the values of the samples are arbitrary. Almost any good teacher of English composition could with a little study and labor prepare a scale that would be just as valid. Each composition is to be rated for Story Value by an analytic-synthetic comparison with the samples on the scale, and then rated for Form Value by counting the number of mistakes in spelling, punctuation, and syntax per hundred words. This scale gives an appearance of accuracy and exactness which the results do not deserve. It is probably a better scale than the average teacher would be able to prepare for herself, but it is lacking in most of the characteristics that would entitle it to serious consideration as a scientific measuring scale.

The Lewis Scales for measuring special types of English composition. — These scales, prepared by Dr. E. E.

Lewis, Superintendent of Schools at Rockford, Illinois, are five in number. One scale measures the quality of Order Letters (ordering goods from a mail-order catalogue), another measures the quality of Letters of Application (for a position), another measures the quality of Narrative Social Letters, another measures the quality of Problematic Social Letters (telling how to do or make something), and the last one measures the quality of Simple Narration. These scales are statistically well built and deserve to be used wherever they will serve more effectively than some previously described form of the Hillegas Scale.

The calculation of percentile values on composition scales. — The best practice in using composition scales is to assign to the child's composition the numerical value of that sample on the scale to which it is most nearly equal in quality. If a number of compositions are assigned to quality 3.8 on the Nassau County Supplement, they may be assumed to lie somewhere between 3.325 and 4.405, — that is, nearer to 3.84 than to either 2.81 or 4.97 in quality. If there are several compositions rated at 5.8 on the Hillegas Scale, it may be assumed that they are distributed evenly over that step on the scale, from down half-way to 4.74, up half-way to 6.75, — that is from 5.30 to 6.30.

Tables LIV and LV are taken directly from the writer's previously published article, " Supplementing the Hillegas Scale." They give the value of each sample and the range of each step to be used in calculating percentile scores from distributions on the Hillegas Scale and on the Nassau County Supplement, respectively. The student will recall that, in locating a given percentile point within a step, one first determines what portion of the total frequencies on

TABLE LIV

To Facilitate the Calculation of Percentile Values in Distributions on the Hillegas Scale for English Composition

Step No.	Value of Sample	Lowest Value	Highest Value	Length of Step	Step No.
0	0.0	0.0	.91	.91	0
1	1.83	.92	2.21	1.30	1
2	2.60	2.22	3.14	.93	2
3	3.69	3.15	4.21	1.07	3
4	4.74	4.22	5.29	1.08	4
5	5.85	5.30	6.30	1.00	5
6	6.75	6.30	7.23	.93	6
7	7.72	7.24	8.05	.81	7
8	8.38	8.05	8.87	.82	8
9	9.37	8.88	?	?	9

TABLE LV

To Facilitate the Calculation of Percentile Values in Distributions on the Nassau County Supplement

Step No.	Value of Sample	Range of Step		Length of Step	Step No.
		Low	High		
0	0.0	0.0	.53	.53	0
1	1.06	.54	1.49	.96	1
2	1.93	1.50	2.36	.87	2
3	2.81	2.37	3.32	.96	3
4	3.84	3.33	4.40	1.08	4
5	4.97	4.41	5.48	1.08	5
6	6.01	5.49	6.61	1.13	6
7	7.22	6.62	7.61	1.00	7
8	8.00	7.62	8.50	.89	8
9	9.0	8.51	?	?	9

that step must be taken, converts this into a fraction of the total range of the step, and adds that fraction of the total length to the lowest possible value in the step. The process is exactly the same with this as with any other scale, but the work is a bit complicated in the case of composition scores, due to the fact that the limits and lengths of the steps are irregular.

To illustrate the use of the above tables in working out percentile values from distributions of scores, consider the following distribution on the Nassau County Supplement.

Score	0	1.1	1.9	2.8	3.8	5.0	6.0	7.2	Total
No. Papers	—	—	3	16	25	4	—	—	48

The fifty percentile or median value will be the value having 24 papers ($48 \div 2 = 24$) above it and 24 below it. After counting the 3 at 1.9, and the 16 at 2.8, one will have to take 5 of the 25 rated at 3.8 in order to count out exactly 24 papers. Taking the lowest 5 of the 25 at 3.8 will mean going 20% (5 is 20% of 25) of the distance from the lowest value of 3.8 to its highest value. From Table LV one may discover at a glance that the step containing sample 3.8 is 1.08 units long. Twenty per cent of 1.08 is .22 of a unit. From Table LV it is also seen that the lowest value of this step is 3.33. Adding the .22 to 3.33 gives 3.55 as the point sought, the median of the distribution.

The finding of any other percentile value is done in the same manner. The first quartile, or twenty-five percentile, for example, would be found by counting out 12 of the 48

papers (25% of 48 = 12). After counting the 3 scored at 1.9, it would be necessary to take 9 of the 16 scored at 2.8 in order to include the desired number of compositions. Taking the lowest 9 of the 16 would mean going .56 of the way through the step (9 is .56 of 16). Table LV shows that this step is .96 units long and begins at 2.37. Taking .56 of .96 gives .54 units, to be added to 2.37. This indicates 2.91 as the point being sought, the twenty-five percentile. Seventy-five per cent of the papers were better, therefore, than quality 2.91 on the scale.

No table will be given here for calculating the percentile scores from distributions on the Thorndike Extension, because it is not probable that many persons will care to use it, except as a scale of reference. Those who do use it may easily calculate for themselves the mid-points between samples, which mark the limits of the steps. One may by subtracting each such mid-point from the next higher mid-point find the length of the step. Those who use the Hudelson Scale may assume that each step is .5 of a unit long and that step 5 reaches from 4.75 to 5.25, that step 5.5 ranges from 5.25 to 5.75, and so on. Those who use one of the Lewis Scales should calculate division points and step ranges before beginning to calculate percentile scores from the distributions obtained.

Summary. — The original Hillegas Scale was built on sound principles and was useful for measuring the general quality of English compositions. The Thorndike Extension is an excellent scale by which to verify doubtful judgments. The Nassau County Supplement is a good scale to employ in introducing teachers to the methods for measuring compositions. Hudelson's scales are the most refined

editions of the Hillegas Scale and should be employed by all judges who have been well trained. Each of these is a scale for measuring quality. Standard test conditions must be maintained if comparisons are to be made as to the relative composition efficiency of different pupils or different classes. The calculation of percentile values from distributions on composition scales requires a careful observation of the statistical range of each score.

QUESTIONS AND EXERCISES

1. Before you undertake to use one of the composition scales, grade in your usual manner, on a 100% scale, a set of compositions written by a class of pupils. Keep on a sheet of paper your record of each child's composition, so there will be no record on the composition paper of what you think it worth. Put away your record and the papers for two months, and then score the entire set in the same manner, without referring to your sheet containing the previous ratings. Make a distribution of the amounts of difference between your first and your second ratings of each paper. Calculate the mean and the mean deviation of these differences.

2. Repeat exactly the same experiment as outlined in No. 1, using some form of the Hillegas Scale for making your ratings in both cases. Be just as careful in your ratings as you were when using your 100% scale. How do the mean and mean deviation of the differences found with the scale compare with those obtained without the scale?

3. Make out a score-card for judging the quality of an English composition. Ask other teachers to make out similar score-cards, allowing 100 points for a perfect composition, and distributing them among the various elements that make up a perfect composition in proportion to the importance of each element. Compare the amounts allowed for each characteristic by the different teachers.

4. Have at least three judges score a set of compositions on one form of the Hillegas Scale, each keeping his records on a sheet of paper. Then have them score the same set of papers in the same manner on at least two other forms of the same scale. Compare the

median scores and the Q's for each of the resulting distributions. Do these comparisons give any indication of which scale it would be best to employ?

5. Rate a set of compositions by comparing them with the scale in one direction only, — beginning at the low end of the scale and going up toward the better samples, for example. Then rate the same compositions by comparing them with the same scale, but comparing them in the opposite direction, — coming down from the good samples to the poorer. How much difference is there between the medians of the scores resulting from these two methods?

6. Plan an experiment that would determine how much difference there would be between the average qualities of compositions resulting from the use of any two composition titles or subjects.

7. Use Table LIV in calculating the twenty-five percentile, median, and seventy-five percentile of each of the following distributions of scores on the Hillegas Scale.

DISTRIBUTION OF SCORES IN GROUP

SCORE	A	B	C	D	E	F	G
9.37	—	—	—	—	—	—	—
8.38	1	—	—	—	1	—	1
7.72	5	4	—	—	4	—	12
6.75	15	16	—	—	12	8	47
5.85	26	22	1	8	31	13	94
4.74	16	10	20	159	13	56	72
3.69	6	1	122	403	1	88	10
2.60	1	—	44	281	—	74	1
1.83	—	—	1	71	—	31	—
0.0	—	—	—	8	—	2	—
Total	70	53	188	930	62	272	237

REFERENCES FOR READING

F. W. Ballou, "Scales for the Measurement of English Compositions," *Harvard-Newton Bulletin*, No. 2, Harvard University.

M. B. Hillegas, "A Scale for the Measurement of Quality in English Composition by Young People," *Teachers College Record*, Vol. 13, No. 4.

E. Hudelson, *Hudelson English Composition Scale.*

E. Hudelson, "English Composition; Its Aims, Methods, and Measurement," Part I, *22nd Yearbook of The National Society for the Study of Education.*

E. E. Lewis, *Scales for Measuring Special Types of English Composition.*

M. R. Trabue, *Supplementing the Hillegas Scale*, Revised edition, Teachers College, Columbia University.

Tests,
To see whether pupils can profit by next step
So the parents may know how he
is studying.
So Supt. can have record in his office
To steer pupils on to better work.
It is customary and custom should
usually be followed.

CHAPTER XVII

MEASUREMENTS OF RELATIONSHIP

EFFORTS to measure the exact amount of ability or information possessed by a pupil or by a class are usually made in the hope that the knowledge thus obtained will be a useful index of other traits or abilities. Intelligence tests are frequently given in the hope that the results may signify how rapidly the pupil will be able to learn new facts and skills. Tests in arithmetic, in writing, in reading, and in spelling are applied in order to determine whether the pupil would probably be able to profit by the next steps in the instructional process. There is usually a looking forward from the results obtained to probable future results in a related field.

The general public tends to accept or to reject a given test on the basis of one or two striking examples of its success or failure. If a phrenologist, for instance, makes several random predictions regarding a child, some one or two of which prove true in a dramatic manner, many people will be willing to consider his methods entirely reliable. The following postal card, received by the writer in April, 1922, from a man who claims to have read character and ability in thousands of faces without having made an error, shows the extravagant claims that can be made where exact methods of measuring relationships are not used.

"CAN YOU TELL

For instance — whether a person is good at figures, or if they can handle figures with ease or have difficulty with even the simplest kind of calculation? The faculty of 'Number' can be easily recognized in the face, and to give you exact knowledge as to how and where to read this faculty you are invited to attend the class lecture

FRIDAY, APRIL 28th, at 8 P.M.

"This will demonstrate how the . . . Method is readily taught in a manner which will enable you to apply the information received in your daily business life."

Occasionally a few statements about the person examined by such an impostor are strikingly true, with the result that hundreds of people go to him every year and pay large fees to have their "faces read." This type of charlatan would not be able to support himself and his practices if the public required exact proof as to the degree of accuracy in his predictions. Likewise, in developing a science of education, no real progress can be made unless teachers and supervisors are willing and able to apply exact measurements to the amount of relationship existing between the results of tests and other measures of success.

The advantages of coefficients of correlation. — A coefficient of correlation is an index figure which summarizes the degree of relationship between two series of measurements of the same individuals. It has much the same character as a measure of average, for it tells in a single two- or three-place number the amount of relationship existing between two rather long series of scores. Without such a mathematical summary, it would often be difficult to form a useful estimate of just how much correspondence exists between different tests or between two trials of the same

test. Suppose, for example, that in two handwriting tests nine pupils made scores on the Starch Scale as follows:

Pupil	Scores in Handwriting Quality	
	First Test	Second Test
A	5	6
B	4	3
C	1	2
D	6	5
E	9	9
F	7	8
G	2	1
H	8	7
I	3	4

It is clear that the above scores do not correspond perfectly, although there is a distinct tendency for those pupils who made low scores in the first test to make similarly low scores in the second. Giving the exact scores for each pupil in the above manner tells how the two measurements agree, or correlate, but it is a crude method of showing the relationship. It would be difficult to tell whether the amount of correlation in the two trials illustrated above was more or less than the amount in two other similarly paired columns of scores. The information that the coefficient of correlation in the two trials listed above was .93, and that in another comparison of results the correlation was .78, is a more easily understood presentation of facts.

By means of such coefficients of correlation one is able to determine which of two intelligence tests gives a more exact

estimate of the relative success of pupils in actual school work. By their aid one may determine which reading test agrees most closely with the known facts about the reading abilities of a class of pupils. They may be used to indicate which of two tests in arithmetic comes nearer giving the same results when repeated a second time; which teacher is most consistent in her marking of pupils' abilities; whether those pupils who work slowly also work more "surely" than those who work rapidly; and how valid are the marks of any particular teacher. These and many other similar questions may be answered by calculating coefficients of correlation.[1]

Easy examples of correlation. — In order to illustrate the methods used in calculating a coefficient of correlation, the student may consider the following scores as though they were the actual results of several different measurements of the same list of pupils. The nine pupils may be assumed to be the same ones listed on the previous page. Their scores are rearranged here in the order of their handwriting ratings in the first test.

The problem in the case of the above examples may be stated as follows: How successful would one be in estimating the scores in Tests B, C, D, E, and F, if he used the scores in Test A as his only basis for making estimates? Or, how are the scores in Test A related to the scores in each

[1] The reader who is not interested in learning to make the calculations need not read the remainder of this chapter. The reader who desires to become expert in these calculations should consult standard books on statistical methods. The presentation of the subject here is very elementary and incomplete, although it will probably be found adequate for the needs of most teachers in the public schools.

TABLE LVI

CORRELATION EXAMPLES: SCORES OF NINE PUPILS IN SIX TESTS

PUPIL	TEST A	TEST B	TEST C	TEST D	TEST E	TEST F
C	1	2	2	3	4	9
G	2	1	4	1	8	8
I	3	4	1	5	1	7
B	4	3	3	9	9	6
A	5	6	5	2	2	5
D	6	5	8	8	7	4
F	7	8	9	4	3	3
H	8	7	6	6	6	2
E	9	9	7	7	5	1

of the other tests? To which is A most closely related, and to which least closely? The only fully satisfactory answer will be a series of coefficients of correlation. The first step in obtaining these coefficients is to build, for each pair of scores, a correlation table showing just what value in Test A goes with each value in each other test. Such correlation tables are given in Figure 43.

Section A of Figure 43 shows how a perfect correlation would look. This might be the correlation in case the same judge rated the same handwriting papers a second time on the same scale. The reader will observe that the marks indicating scores run in a straight line from the lower left-hand corner to the upper right-hand corner of the diagram. Section B shows the correlation previously described as having a coefficient of .93, being the correlation between the supposed first and second tests of handwriting quality. It will be noticed that there is not a simple straight line in

this case, although the drift of the scores is distinctly in the same direction, indicating that an increase of ability in one test is accompanied in a general way by an increase of ability in the other. It is easier to understand the degree of relationship indicated by a coefficient of .93 from an examination of Section B of Figure 43, than from examining columns A and B of Table LVI.

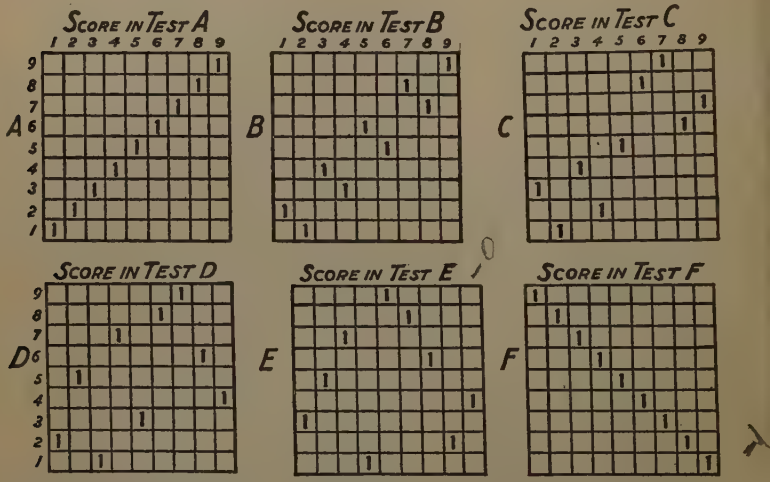


FIG. 43. — Correlation tables showing various degrees of relationship between scores in Test A and scores in other tests: $A, r = 1.00$; $B, r = .93$; $C, r = .78$; $D, r = .47$; $E, r = 0$; $F, r = -1.00$.

Section C, representing the relation of Test A to Test C, shows a tendency to drift off in the same general direction followed in Section A, but the scores are not so uniformly near the direct relation line. Section D, which shows the relationship between Test A and Test D, shows a still greater departure of the scores from a perfect relationship. From the appearance of Section E one would judge that there was

no direct relation whatever between scores in Test A and scores in Test E. Section F, however, shows a perfect *negative* relationship, in which every gain in Test A is accompanied by a corresponding loss in Test F.

For many comparisons the average teacher will obtain more helpful information by plotting correlation tables like those in Figure 43 than by working out the coefficient of correlation. It is frequently desirable, however, to be able to express the amount of relationship in a single index figure, which may be compared with other index figures expressing similar relationships. The formula used is, $r = \frac{\Sigma x \cdot y}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$, in which the x's are the deviations of the measures in the first series from their average, and the y's are the deviations of the measures in the second series from their average. The letter r stands for the coefficient of correlation. The formula reads as follows: The coefficient of correlation is the quotient of the algebraic sum of the products of each x deviation and its corresponding y deviation, divided by the product of the square root of the sum of the squares of the x deviations multiplied by the square root of the sum of the squares of the y deviations.

To illustrate the use of the formula, consider the relation between Test A and Test D, as illustrated in Section D of Figure 43, and repeated in Table LVII. The column headed x, immediately to the right of the blocks, shows how much each of the scores in Test A deviates from their average, a score of 5. The line of numbers following y, just below the list of scores in Test D, shows how much and in which direction each score in D deviates from the average, which is also 5. One horizontal and one vertical row of

$r = \frac{\Sigma xy}{N \sigma_x \sigma_y}$

squares have been inclosed in heavy lines to show that the deviations in these rows are zero, and to aid the eye in making calculations.

TABLE LVII

CORRELATION BETWEEN TESTS A AND D

Score in Test A	−4	−3	−2	−1	0	1	2	3	4	x	$x \cdot y$	x^2
9							1			4	8	16
8						1				3	3	9
7				1						2	−2	4
6								1		1	3	1
5		1								0	0	0
4									1	−1	−4	1
3					1					−2	0	4
2	1									−3	12	9
1			1							−4	8	16
Score in Test D	1	2	3	4	5	6	7	8	9		34−6	60 = Σx^2
y	−4	−3	−2	−1	0	1	2	3	4	$\Sigma x \cdot y = 28$		
y^2	16	9	4	1	0	1	4	9	16	= 60 = Σy^2		

The squares of the deviations, called for in the denominator of the correlation formula, are given in the column at the extreme right, totaling 60 for the x squares, and in a row at the bottom, giving again a total of 60 for the y squares. The products of the x and y deviations, called for in the numerator of the formula, are given between the column of x's and the column of x squares. Beginning at the top, it will be observed that one pupil had a deviation of 4 in x and 2 in y, which makes his $x \cdot y$ product 8. The next pupil had a deviation of 3 in x and 1 in y, giving him 3 as a product. The next had 2 in x and -1 in y, giving him -2 as a product. At the bottom of this column the algebraic sum of these deviation products is found to be 28, which therefore becomes the numerator of the fraction in the formula.

Substituting these values in the formula $r = \frac{\Sigma x \cdot y}{\sqrt{\Sigma x^2}\ \sqrt{\Sigma y^2}}$ gives, $r = \frac{28}{\sqrt{60}\ \sqrt{60}}$, which when the radicals are removed from the denominator gives, $r = \frac{28}{60} = .47$, or more nearly that than any other two-place decimal. The correlation between Test A and Test D, as far as the scores of these nine pupils may be taken as an indication of it, is therefore .47, which is about the amount of correlation one would normally obtain in a class of forty pupils between reading test scores and intelligence test scores.

TABLE LVIII

CORRELATION BETWEEN TESTS A AND C

TEST C / Test A	1	2	3	4	5	6	7	8	9	x	$x \cdot y$	x^2
9							1			4	8	16
8						1				3	3	9
7									1	2	8	4
6								1		1	3	1
5					1					0	0	0
4			1							−1	2	1
3	1									−2	8	4
2				1						−3	3	9
1		1								−4	12	16
y	−4	−3	−2	−1	0	1	2	3	4	$\Sigma x \cdot y = 47$		$\Sigma x^2 = 60$
y^2	16	9	4	1	0	1	4	9	16	$\Sigma y^2 = 60$		

$$r = \frac{47}{\sqrt{60}\ \sqrt{60}} = \frac{47}{60} = .78$$

As a further illustration of the use of the formula take the relation between Test A and Test C, illustrated in Section C of Figure 43, and repeated here in Table LVIII. The average in both sets of scores is again 5, so the inclosed

lines of the table, indicating deviations of zero, are at the same place as in the previous example. The sums of the squares of the deviations are 60, as before. The first $x \cdot y$ deviation product, at the top of the page, is 4 in A and 2 in C, making 8 as a product. The next is 3 in A and 1 in C, giving 3 as a product. There are no negative products, since all scores fall in the $--$ and $++$ quadrants of the table. The sum of the deviation products in this table is 47, which is therefore the numerator of the correlation fraction. The denominator is again 60, making the final coefficient .78, which is a little smaller than the coefficient of correlation to be expected when the same test is used a second time on a class of pupils.

One more example may be considered. The appearance of Section E of Figure 43 does not indicate any definite relationship at all. The amount of A does not seem to be any indication of the amount of E that might be expected. Table LIX gives the solution for the coefficient of correlation.

Again the squares of the deviations equal 60, when added. The first $x \cdot y$ deviation product at the top of the diagram is 0, because 4 in Test A multiplied by 0 in Test E gives 0. Three in A multiplied by 1 in E gives 3 as a product. Minus two in A multiplied by 2 in E gives -4 as a product. The final sum of the deviation products is $17 - 17$, which equals zero. Zero divided by 60, or by any other amount, would be zero. The coefficient of correlation between Test A and Test E was therefore 0 for this group of nine pupils. There was no relation between the two sets of scores. There was not even a slight tendency for the scores to drift in the same direction. They scattered all over the diagram.

TABLE LIX

CORRELATION BETWEEN TESTS A AND E

TEST E	1	2	3	4	5	6	7	8	9	x	$x \cdot y$	x^2
9					1					4	0	16
8						1				3	3	9
7			1							2	−4	4
6							1			1	2	1
5		1								0	0	0
4									1	−1	−4	1
3	1									−2	8	4
2								1		−3	−9	9
1				1						−4	4	16
y	−4	−3	−2	−1	0	1	2	3	4		17−17	60
y^2	16	9	4	1	0	1	4	9	16	$x \cdot y = 0$		$x^2 = 60$

Test A

$y^2 = 60$

$$r = \frac{0}{\sqrt{60}\ \sqrt{60}} = 0$$

The student may work out for himself the coefficients in the other cases shown in Figure 43. Section A, showing a perfect relationship, will give a sum of deviation products amounting to 60. When this sum is divided by 60, the denominator in these examples, the coefficient obtained as a quotient will be 1.00, which indicates perfect correlation. The sum of the deviation products in Section F will be −60, which will give −1.00 as a coefficient of correlation. In Section F, therefore, one may see an example of perfect negative correlation, — a case where an increase in score in one test means a corresponding and proportional loss in score in another test. Coefficients of correlation may range in size all the way from 1.00 down through 0 to −1.00, although the majority of those worked out in educational fields range from .90 down to .20, or thereabouts.

Ordinary examples of correlation. — The examples given above differ from those ordinarily met in that no two pupils were assigned the same score in the same test, while in actual school tests there are strong probabilities that several pupils will make equal scores in some of the tests. The artificial scores are easier to use in learning the method, however.

The writer gave an intelligence test and then a reading test to a class of more than fifty students of educational measurements. Some member of the class raised the question as to whether or not the two were not very similar in the abilities they measured. In order to answer the question, a coefficient of correlation was worked out between the two sets of scores. The scores of the first nineteen students, taken alphabetically, were as follows:

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Intelligence	7	13	16	12	13	14	11	8	14	15	10	12	13	10	9	13	9	16	10
Reading	6	12	19	10	13	12	12	7	16	11	13	12	12	11	13	14	6	20	10

These scores may be tabulated in a correlation table as shown in Table LX.

The first column at the right of the tabulations is the frequency distribution for the intelligence scores, and the first row at the bottom the frequencies for the reading scores. Each of these two lists has a total of 19 scores. The median of the intelligence distribution is 12.75, and the mean is 12.34, so one may take the score of 12(12.0 — 13.0) as the intelligence score from which to calculate deviations. The median reading score is 12.5, and the mean 12.55, so a score of 12 may be used as the point of reference in this case also. These scores have therefore been inclosed in

TABLE LX

A Correlation between Intelligence and Reading

Scores in Reading Test

Intelligence Scores	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	f	x	$f \cdot x \cdot y$	$f \cdot x^2$
16														1	1	2	4	60	32
15						1										1	3	−3	9
14							1				1					2	2	8	8
13							2	1	1							4	1	3	4
12					1		1									2	0	0	0
11							1									1	−1	0	1
10					1	1		1								3	−2	4	12
9	1							1								2	−3	15	18
8		1														1	−4	20	16
7	1															1	−5	30	25
f (frequency)	2	1			2	2	5	3	1		1			1	1	19		140−3	125
y	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8			= 137	Σfx^2
$f \cdot y^2$	72	25			8	2		3	4		16			49	64	$\Sigma f \cdot y^2 = 243$			

$$r = \frac{x \cdot y}{\sqrt{x^2}\ \sqrt{y^2}} = \frac{137}{\sqrt{125}\ \sqrt{243}} = \frac{137}{11.18 \times 15.59} = \frac{137}{174.29} = .79$$

heavy lines to show that they furnish the zero deviations. The second column at the right and the second at the bottom indicate the amount by which each score differs from the average.

The calculations in this case differ from those in the simple examples chiefly in that the frequency of each score must be considered. In obtaining the sums of the y squares, for example, it is necessary to note that there are two individuals with a y deviation of -6. The square of -6 is 36, which when multiplied by 2 gives 72. Seventy-two is therefore recorded in the appropriate $f \cdot y^2$ row below the y deviations. There is only one deviation of 5, so one may write 25 in the $f \cdot y^2$ row. The column at the extreme right is made up in the same manner. The fact that two stu-

dents have a deviation of 4 in intelligence makes 32 the square of their amount of x deviation ($2 \times 4^2 = 32$). The total of the $f \cdot x^2$ deviations is 125, and the total of the $f \cdot y^2$ deviations is 243.

Beginning at the top of the table, one may calculate the $f \cdot x \cdot y$ products as follows: one deviation of 8 in y multiplied by 4 in x gives 32, to which must be added another of 7 in y multiplied by 4 in x (a product of 28). The algebraic sum of these two is 60, which is therefore recorded at the top of the column headed $f \cdot x \cdot y$. The next case has a y deviation of -1 and an x deviation of 3, making the product -3. In some instances there are both positive and negative products, in which case their algebraic sum is recorded under $f \cdot x \cdot y$. The three students who made 10 in the intelligence test, for example, have deviation products of 4, 2, and -2; the algebraic sum of which is 4, as recorded in the appropriate column. The sum of all $f \cdot x \cdot y$ products is $140 - 3$, or 137. The square root of the x squares is 11.2, and the square root of the y squares is 15.6. The product of these is 174.3, which is the divisor in the correlation fraction, thus giving .79 as the final coefficient of correlation. This amount of relationship between intelligence and reading ability is unusually high.

The number of individuals included in a comparison has a large influence on the reliability of the correlation coefficient obtained. On the whole it is desirable to calculate a coefficient with a "population" of at least fifty individuals, although a table given later in this chapter indicates the amount of reliability for coefficients of various sizes and with populations of varying size. Table LXI has been prepared to show the type of correlation problem most

commonly met. The scores are the scholarship and conduct marks assigned by grammar-grade teachers to fifty pupils in a private school in Philadelphia. The pairs of marks from which the table was made were as follows:

Sch.	Cond.	Sch.	Cond.	Sch.	Cond.	Sch.	Cond.	Sch.	Cond.
71	91	74	92	75	97	70	69	83	94
71	87	79	81	78	70	71	78	76	89
69	77	78	80	69	82	81	90	66	87
79	85	68	55	73	81	64	87	70	87
76	94	83	87	76	88	61	80	80	88
56	90	69	94	87	94	63	92	77	59
76	72	70	83	75	82	69	72	54	86
76	81	76	92	68	86	56	88	70	86
88	98	78	93	72	87	77	78	70	82
72	87	70	92	69	86	75	92	80	95

The reader will observe that the grades have been grouped in twos, 54 with 55, 56 with 57, and so on. This is entirely legitimate, and is often desirable, in order to reduce the number of groups in each measurement to approximately twenty. Having more than twenty groups adds very little to the accuracy of the work and makes a larger number of opportunities for errors in computation. The line and column giving the amounts of deviation from the average have been omitted, since putting them in makes unnecessary labor when their squares are also recorded. It will be observed, both from the appearance of the correlation table and the size of the coefficient of correlation, that there is little relation between the teachers' estimates of scholarship and their estimates of conduct in this grammar school.

TABLE LXI

A Correlation between Scholarship and Conduct Marks

Marks for Conduct in School

Scholarship Marks	54 55	56 57	58 59	60 61	62 63	64 65	66 67	68 69	70 71	72 73	74 75	76 77	78 79	80 81	82 83	84 85	86 87	88 89	90 91	92 93	94 95	96 97	98 99	f	$f \cdot x$	$f \cdot x \cdot y$	$x^2 f \cdot x^2$
88–																							1	1	48	64	64
86–																					1			1	28	49	49
84–																										36	
82–																	1				1			2	20	25	50
80–																		1	1		1			3	28	16	48
78–									1					2		1				1				5	−36	9	45
76–			1							1			1	1						1	1			8	−38	4	32
74–															1					2		1		4	9	1	4
72–														1			2							3	0	0	0
70–								1				1		2			3		1	1				9	12	1	9
68–	1								1			1		1			2				1			7	52	4	28
66–																	1							1	0	9	9
64–																	1							1	0	16	16
62–																					1			1	−15	25	25
60–														1										1	18	36	36
58–																									0	49	
56–																			1	1				2	−24	64	128
54–																	1							1	0	81	81
f	1		1						1	1	2		1	2	5	4	1	1	4	3	6	5	1	1	50	215	−113	642
y^2	256		196						81	64	49		25	16	9	4	1	0	1	4	9	16	25	36		=102		
$f \cdot x^2$	256		196						81	64	98		25	32	45	16	1		4	12	54	80	25	36		$\Sigma f \cdot y^2 = 1025$		

$$r = \frac{102}{\sqrt{624}\ \sqrt{1025}} = \frac{102}{24.98 \times 32.01} = \frac{102}{799.6} = .128, \text{ or approximately } .13$$

The reliability of a coefficient of correlation. — Any coefficient actually obtained is, of course, from only a small sample of the total population. The true correlation for all persons can never be calculated because it will never be possible to measure all people. There is a possibility of calculating mathematically, however, from any obtained coefficient approximately what the chances are that the true coefficient would differ by any given amount from the one obtained. The formula used for this purpose is $\sigma = \frac{1 - r^2}{\sqrt{n}}$, in which σ indicates the standard deviation of

the difference between the true coefficient and that which has actually been obtained, and n indicates the number of individuals used in obtaining the correlation, r.

Since the median deviation is a more easily understood measure of variability and is usually .6745 of σ, the above formula may be written, P. E. $= .6745 \frac{1 - r^2}{\sqrt{n}}$. Table LXII has been prepared by this formula to give approximate values for the P. E. of a coefficient of correlation of various sizes, computed from populations of varying size.

TABLE LXII

MEDIAN DEVIATION OF THE DIFFERENCES BETWEEN TRUE AND OBTAINED CORRELATIONS [1]

FOR AN r OF	NUMBER OF INDIVIDUALS IN COMPARISON										
	20	25	30	35	40	45	50	60	70	80	90	100
.10	.15	.13	.12	.11	.11	.10	.09	.09	.08	.08	.07	.07
.20	.14	.13	.12	.11	.10	.10	.09	.08	.08	.07	.07	.06
.30	.14	.12	.11	.10	.10	.09	.08	.08	.07	.07	.06	.06
.40	.13	.11	.10	.10	.09	.08	.08	.07	.07	.06	.06	.06
.50	.11	.10	.09	.09	.08	.08	.07	.07	.06	.06	.05	.05
.60	.10	.09	.08	.07	.07	.06	.06	.06	.05	.05	.05	.04
.70	.08	.07	.06	.06	.05	.05	.05	.05	.04	.04	.04	.03
.80	.05	.05	.04	.04	.04	.04	.03	.03	.03	.03	.03	.02
.90	.03	.03	.02	.02	.02	.02	.02	.02	.02	.01	.01	.01

The formula and Table LXII give the median deviation of the differences between an obtained coefficient of corre-

[1] For more exact values the reader may substitute the value obtained for r and the number of individuals, n, in the formula given in the text.

lation and the coefficient that would probably be obtained if one were able to measure every individual and to calculate a correlation coefficient from the resulting measures. Suppose, for example, that one obtains a coefficient of .70 between two traits, from a correlation of the measures of 45 individuals. The table indicates that the median deviation in this case would be .05. The result may be written, $r = .70$, P. E. .05, or $r = .70 \pm .05$, which means that if an infinite number of such coefficients were worked out, half of them would be between .65 and .75. In other words, the chances are even that the true coefficient does not differ from that obtained by more than .05.

Interpretation of coefficients of correlation. — The only certain fact that one can infer from a high coefficient of correlation is that the two traits are so related that high ability in one is associated with high ability in the other. It is not safe to infer, because the two have a high coefficient of correlation, that one measure is the result of the other.

Methods of multiple and partial correlation have been developed, which enable one to determine how much relation exists between any two measured characteristics when any other measured trait or combination of traits has been rendered constant. The work involved in this type of calculation is too complex for presentation in an elementary book on educational measurements, but scientific educators must learn the technique and use it frequently if the maximum of progress is to be made in educational practice. By its means such questions as the following may be answered: " Exclusive of general intellectual ability, how much relation is there between ability to recognize and classify words and the ability to read and

obey instructions rapidly? " This factor of general intellectual strength and maturity frequently enters into correlations between different school subjects, making them yield a much higher correlation coefficient than their direct relation to each other would warrant. The ability to calculate relationships, uninfluenced by such factors, is worth developing.

One of the questions often asked refers to the size of coefficient necessary to indicate a really significant relationship. There is no point between 0 and 1.00 at which a coefficient may be said to become significant. In some cases a correlation of .30 might be considered remarkably high, while in other cases it would be too low to be assigned any weight whatever. The significance of the coefficient depends upon the character of the measures being compared as well as on its own size. The variability of the group measured is a determining factor, making the coefficient larger when the dispersion of the measures is large, and smaller when the deviations are small. Two forms of the same reading scale, for example, might have a correlation of .90 if the pupils measured came from all grades from the 3d to the 8th, while in any one grade the coefficient might not be higher than .60. In quoting coefficients of correlation one should therefore describe the group of individuals measured as well as state the size of the result.

Correlations as measures of reliability. — One of the purposes for which the coefficient of correlation may well be used is the determination of the reliability of any given test. How well does a test correlate with itself? If two forms of the same test have a low correlation with each other, one need not expect either of them to correlate well with any other ability. Unless two measurements with

the same instrument give a correlation of at least .75 with each other, the test can have relatively little value in educational work. The reliability of such a test might often be improved, however, by making it longer, or by giving several additional forms of it.

A formula has been devised for determining how much lengthening would be required in order to obtain any given reliability coefficient. This formula may be stated as follows: $N = \frac{r_x(r_1 - 1)}{r_1(r_x - 1)}$, in which N is the number of times their present length the forms must be made in order to give a coefficient of r_x, when the actual correlation coefficient between two forms of the present length is r_1. Suppose, for example, that two forms of a test give a correlation of .72, and it is desired to get a correlation of .95 between lengthened forms. To find how many times their present length the tests will need to be, one may substitute in the above formula as follows:

$$N = \frac{.95(.72 - 1.0)}{.72(.95 - 1.0)} = \frac{.95(-.28)}{.72(-.05)} = \frac{.266}{.036} = 7.2.$$

The test forms would have to be 7.2 times their present length in order to increase the reliability from .72 to .95.

This formula is usually given in the other form, $r_x = \frac{Nr_1}{1 + (N - 1)r_1}$, in order to enable one to calculate from any obtained coefficient just how reliable the test would be if repeated or made a given number of times its present length. The only difference is that in one case the solution is for N, and in the other it is for r_x. In either statement of the formula, r_1 is the known quantity and is

obtained from a correlation of the two halves or of two different forms of the test.

Summary. — The coefficient of correlation is an index figure summarizing the relationships between two measurements of the same group of individuals in much the same manner as an average summarizes a distribution of scores. Making a correlation table is the first step in calculating a coefficient and is in itself a good visual representation of the relationship being measured. A high coefficient of correlation indicates that a high score in one test is usually accompanied by a correspondingly high score in the other. A low correlation indicates that the score in each test varies from its average without much relation to the variations in the other test. The self-correlation of a test is employed as a measure of its reliability.

QUESTIONS AND EXERCISES

1. Seven pupils were measured in three different traits, with scores as follows:

Test A	Test B	Test C
1	2	6
2	4	12
3	6	4
4	8	14
5	10	2
6	12	10
7	14	8

Prepare correlation tables and calculate the relationship between Test A and each of the other two.

2. Examine the table on the following page. Which of the six tests correlates most closely with the teacher's estimate of her thirty-three pupils? What is the coefficient of correlation in each case?

Scores of a Sixth Grade

Pupil No.	General Teacher's Estimate	Arithmetic			Language		Spelling
		Multiplication		Reasoning	Scale A (*d*)	Composition (*e*)	(*f*)
		Woody (*a*)	Courtis (*b*)	Stone (*c*)			
1	14	22	6	4	14	35	9
2	19	16	7	4	18	35	10
3	1	19	0	1	11	25	2
4	15	14	7	3	17	35	6
5	30	24	8	5	28	50	13
6	16	22	8	4	23	40	8
7	2	17	6	1	19	35	10
8	13	20	7	3	14	35	7
9	8	21	6	3	18	30	3
10	7	21	9	3	14	35	2
11	25	20	5	4	23	40	14
12	4	18	8	5	19	40	8
13	6	23	5	4	14	35	7
14	28	25	6	7	20	40	11
15	29	25	3	7	20	50	8
16	23	22	8	6	19	50	4
17	32	27	4	5	24	60	16
18	9	22	10	3	14	35	9
19	31	25	7	8	25	30	4
20	27	23	6	7	22	35	9
21	3	15	10	0	15	35	4
22	5	22	5	4	15	35	9
23	20	23	11	9	23	45	10
24	26	20	6	5	22	35	15
25	21	26	6	5	25	40	13
26	12	22	8	2	16	40	6
27	18	26	9	3	20	35	13
28	17	16	5	2	13	30	8
29	11	16	2	0	16	40	4
30	10	23	6	4	17	35	8
31	24	22	6	3	15	40	8
32	22	20	10	1	20	35	7
33	33	22	4	6	30	40	13

3. What is the coefficient of correlation between the scores in column I and the scores in each of the other two columns?

Individual	I Sums per Minute	II Minutes per Sum	III Sums in 5 Minutes
1	6	.166	30
2	15	.066	75
3	11	.091	55
4	3	.333	15
5	5	.2	25
6	9	.111	45
7	20	.05	100
8	7	.143	35
9	9	.111	45
10	1	1.0	5
11	18	.055	90
12	4	.25	20

4. What is the median deviation (P. E.) of each of the coefficients found in question 3?

REFERENCES FOR READING

C. A. Gregory, *Fundamentals of Educational Measurement*, Chapter XII.

D. C. Jones, *A First Course in Statistics*, Chapters X and XI.

T. L. Kelley, *Statistical Method*, Chapter VIII.

W. I. King, *Elements of Statistical Method*, Chapter XVII.

H. O. Rugg, *Statistical Methods Applied to Education*, Chapter IX.

E. L. Thorndike, *Mental and Social Measurements*, Teachers College, Chapter XI.

CHAPTER XVIII

TESTS OF INTELLECTUAL ABILITY

NUMERICAL statements of the amount of intellectual ability possessed by a child or by an adult have been possible during the past few years, although relatively little use has been made of such measurements. As larger and larger numbers of teachers and other officials make use of them, scales and tests for obtaining mental scores will undoubtedly continue to be improved. Even now they are sufficiently well developed, so that no progressive school can afford to do without them. Present instruments for measuring intellectual strength are in about the same stage of refinement as are the instruments for measuring school achievements. A few have been well constructed and well standardized, while others are relatively unsatisfactory in construction or in standardization.

The first objective intelligence tests. — The general public has usually assumed that there are only two grades of intellect, the normal and the abnormal. It has been quite customary for the average person to classify minds as either sound or unsound. In most instances those of sound minds have been further classified on some such crude scale as Dull-Average-Bright-Brilliant, or Slow-Average-Quick. The difficulty with these classifications is the one which is common to all subjective scales: the very same mind would be called "average" by one person, "dull" by another, and "bright" by a third judge. The

need for some definite standard of comparison, which all could know and understand, is shown by the wide use of such qualifying phrases as " sharp as a tack," and " slow as molasses in winter.". Although picturesque and interesting, such analogies are too inaccurate to serve any really useful purpose.

The real beginnings of modern objective tests of intelligence are probably to be found in the work of Sir Francis Galton on the differences between the mental abilities of different individuals. The work of Galton in England was supplemented and refined in America by Professor J. McKeen Cattell and other psychologists, whose tests were chiefly of the differences in the sense organs of different individuals. It was assumed that greater intelligence was associated with and indicated by greater sensitivity to pain, quicker reaction time, greater auditory acuity, keener perception of differences in weight, longer memory span, and the like. Although these tests required a considerable amount of special apparatus, psychological laboratories in all parts of the country were quick to adopt them.

The Binet-Simon Tests. — While the above tests for individual differences were being perfected, a French psychologist, Alfred Binet, was developing a more systematic series of easily administered tests for the detection of low mental ability. In coöperation with Dr. T. Simon, a French physician, Binet selected a series of thirty simple tasks, each of which seemed to require for its solution a certain minimum amount of mental strength and maturity. This series of tests, published in 1905,[1] was accompanied

[1] "New Methods for the Diagnosis of the Intellectual Level of Subnormals," *L'Année Psychologique*, 1905, pp. 191–244.

by suggestions for the medical and anatomical examination of persons suspected of having low intellectual abilities. The psychological series included such tasks as pointing to different parts of the body, counting coins, obeying directions, naming the days of the week, giving the names of familiar objects, and discriminating between small blocks having different weights. The elements of intelligence which Binet sought to measure were "judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances."

Three years later, Binet published a revision and extension of his psychological examination,[1] based upon intensive study of a considerable number of normal and feeble-minded children. The number of tasks was increased to fifty-six in the 1908 revision, and they were grouped by ages, from three to thirteen, according to the age a typical child would normally be when he could first perform each task successfully. A still further revision of the tests was published by Binet in 1911,[2] together with data regarding the relations discovered between scores in these tests and such measures as school progress, social status, and teachers' judgments.

Dr. Henry Goddard probably deserves the greatest amount of credit for acquainting American psychologists

[1] "The Development of Intelligence in the Child," *L'Année Psychologique*, 1908, pp. 1–90.

[2] "New Investigations upon the Measure of the Intellectual Level among School Children," *L'Année Psychologique*, 1911, pp. 145–201. English translations of the articles by Binet are given in a volume published by the Training School, Vineland, N. J., entitled, *The Development of Intelligence in Children*. This volume is of unusual value to any one interested in this work.

and educators with the Binet Tests, although other workers were quick to recognize the value of the method and to experiment with it. The use of these tests in classifying four hundred feeble-minded children in the Training School and their further application to almost two thousand children in the public schools at Vineland, New Jersey, may be looked upon as the chief source of inspiration for the extensive checking and modification to which the tests have since been subjected.[1]

Binet mental ages and intelligence quotients. — One of the elements which contributed to the popularity of the Binet Tests was the statement of their resulting scores in terms of years and months. The idea that a child's mind need not necessarily have the same degree of maturity as his body was not wholly new, but the possibility of comparing " mental age " and " chronological age " was novel, and actual use of such comparisons proved to be unexpectedly interesting and helpful.[2] Many of those who have employed the Binet Tests in America have been more confident of the accuracy of the mental age scores than was their originator, who recognized the dangers that accompany his easily understood method of indicating mental power.[3]

[1] H. H. Goddard, "Four Hundred Feeble-minded Children Classified by the Binet Method," *Pedagogical Seminary*, 17, pp. 387–397.

H. H. Goddard, "Two Thousand Children Measured by the Binet Measuring Scale of Intelligence," *Pedagogical Seminary*, 18, pp. 232–259.

[2] "Mental age" is the age index resulting from the application of the Binet Tests.

"Chronological age" is the actual age of the child, found by counting the number of years, months, and days since his birth.

[3] In the 1908 article, previously mentioned, Binet wrote as follows:

Over-confidence in the accuracy of mental age scores has been especially noticeable since the "quotient" technique has been employed in connection with such scores. The "intelligence quotient," or "I. Q.," is the name that has been applied to the ratio obtained by dividing the mental age by the chronological age. If the tests were accurately gauged, a child who was perfectly normal in his mental development would at ten years of age show a mental age of ten years, which would result in an intelligence quotient of 100 $\left(\frac{10}{10} = 1.00\right)$. Similarly, the normal twelve-year-old child would possess an I. Q. of 100, provided the tests were accurately adjusted to render a correct twelve-year mental age. If some bright ten-year-old child was found by the tests to have a mental age of twelve years, his intelligence quotient would be 120 $\left(\frac{12}{10} = 1.20\right)$. On the other hand, a ten-year-old child with a mental age of eight years would have an intelligence quotient of 80 $\left(\frac{8}{10} = .80\right)$.

The use of intelligence quotients was stimulated by the discovery that in most cases the I. Q. remains almost constant as the child grows older. The six-year-old child who has a mental age of eight years, and therefore an I. Q. of 133, will usually be found at ten years of age to have a mental age of about thirteen and three-tenths years. And on the other hand, a six-year-old child with a mental age of

"A child is backward in certain of the tests for his age and in advance for others. We think it is possible to overcome this difficulty; but it is on condition that we adopt some arbitrary rule; and the said rule, however good it may be, will always give an artificial character to the procedure, and to such a degree that if, by chance, one had adopted another rule, one would arrive at quite different results."

five years will probably show a mental age of ten years when he reaches a chronological age of twelve years. Although this ratio does not remain absolutely constant as the child develops, it is so nearly constant that one may from his intelligence quotient predict with reasonable accuracy what a child's mental age will be at any point in his career. Because of this rather startling relationship between mental and chronological age, some Binet testers have tended to assume that there is a supernatural infallibility in mental-age scores and their derived quotients. One should bear in mind constantly, however, that slight errors in the tests or in their application may partially invalidate both the mental-age scores and the derived quotients.

Revisions and modifications of the Binet Tests. — Those interested in intelligence tests soon realized that the mere word for word translation of the Binet Tests was not sufficient. Numerous " versions " of the Binet Tests were therefore used in various centers throughout the United States, and many proposals were made for their improvement and extension. The Yerkes-Bridges Point Scale was perhaps the most important of the early modifications.[1] The Point Scale eliminated the age grouping of tests and the accompanying mental-age scores. It also gave partial credits for incomplete solutions of certain test problems, while in the Binet series each task must be correctly done before any credit may be allowed. In general, it may be said that the point-scale method did not prove to be especially superior to the mental-age method.

Because of the extent to which success in the Binet Tests

[1] R. M. Yerkes, J. W. Bridges, and R. S. Hardwick, *A Point Scale for Measuring Mental Ability.*

depends upon the ability to understand and carry out oral directions, and because of the relatively large amount of language content in the test questions themselves, several attempts have been made to develop "non-language" tests for intelligence. The Binet series is not fully satisfactory for the measurement of those who do not understand the English language. The most widely-known of the non-language series of tests is the Pintner-Paterson Scale of Performance Tests.[1] In the performance tests, pupils work with pictures, blocks, and other objects, rather than with words and their meanings. The instructions in the Pintner-Paterson Tests involve a minimum of spoken words, being given chiefly by signs and gestures. Success in the performance tests is not perfectly correlated with success in the Binet Tests, but there seems to be no reason for denying the claim that performance tests measure certain phases of the group of characteristics usually called intelligence.

The most widely-recognized revision of the Binet Tests was made by Professor Lewis M. Terman and his associates at Leland Stanford University in California. The Stanford Revision of the Binet Scale is thus far the standard against which other intelligence tests are most frequently checked. Dr. Terman found it desirable not only to transfer certain test questions from one age-group to another but also to add a great many new test elements and to modify many of those that were retained. These changes were based on a most painstaking study of the results obtained by testing hundreds of normal children in the public schools as well as many defective and superior children. The series

[1] R. Pintner and D. G. Paterson, *A Scale of Performance Tests.*

was also extended upward by standardizing new tests on various groups of adults. The Stanford Revision of the Binet Scale is therefore designed to measure intelligence in all grades of individuals from the typical three-year-old child up to the superior adult.[1]

A more recent revision of the Binet Tests is that made by Dr. John P. Herring.[2] The materials used are similar to but not identical with those used in the Binet Tests, and the method of administering it is similar to that used by Binet. The scoring, however, is quite different. Instead of allowing two months mental-age credit for each successful response, which is essentially the method used in the Stanford Revision, numerical credits are allowed for each response. The sum of all these credits constitutes the "total numerical score" in the test. Convenient tables are supplied by which one may convert any total numerical score into a mental-age score, thus making the results of Herring's Revision equivalent for all practical purposes to the results of the Stanford Revision.[3]

Group tests of mental ability. — In all of the modifications of the Binet Tests, an examiner must spend from fifteen minutes to an hour examining each individual child. In the course of a single school day it is possible for an efficient Binet examiner to test eight or ten pupils. To measure by

[1] L. M. Terman, *The Measurement of Intelligence*, especially Chapter V. See also L. M. Terman, *The Stanford Revision and Extension of the Binet-Simon Measuring Scale of Intelligence.*

[2] J. P. Herring, *Herring Revision of the Binet-Simon Tests.*

[3] Herring reports that the coefficient of correlation between the Stanford Revision and his own is between .95 and .99 when unselected age-groups are tested.

this method the intelligence of a class of forty pupils would therefore require almost a week, unless a considerable part of the testing were done outside school hours. It is quite evident that such a program is more expensive than many schools would care to undertake, except after confidence in the advantages of such tests had been thoroughly established.

The introduction of the Binet Tests in America occurred at about the same time as the early development of educational tests by Thorndike and others.[1] The educational tests were from the very first designed to measure pupils in groups and classes rather than individually. The advantage of group examinations over individual examinations in mere economy of time was so obvious that attempts were soon made by various investigators to develop group tests of intelligence.[2]

Dr. Arthur S. Otis, working under the direction of Professor Terman at Leland Stanford University, was the first to develop a short yet comprehensive series of group tests for measuring intelligence. The chief contribution of the Otis Tests lies in their convenient arrangement for checking errors in the pupil's work. The test questions or problems are printed on the page in such a manner that the child's answer must appear in a specified place. In most cases the answer is to be indicated by underlining or checking

[1] See Chapter III.

[2] It was with this purpose in view that the writer modified and standardized the language completion test for use in schools. M. R. Trabue, *Completion Test Language Scales*, Teachers College, Contributions to Education, No. 77. Other workers undertook to standardize the opposites tests, the mixed relations tests, and the like.

some printed symbol on the page rather than by writing anything new. Transparent " stencils " are provided for scoring the Otis Tests. These stencils when superimposed on the test page indicate exactly where each check-mark should be, thereby making it possible to count the errors on a page in a brief fraction of a minute and without stopping to read any part of any test question. With the introduction of these mechanical improvements in the tests, the possibility of employing relatively inexpensive clerical workers to score the papers greatly reduced the cost and increased the popularity of intelligence testing.

Army intelligence tests. — When the United States was drawn into the World War in 1917, its army was so small as to be almost insignificant in comparison with those already fighting in Europe. There were millions of young men available, but neither they nor any one else had an adequate idea of where or how to fit them into the organizations necessary to carry on successful warfare. The psychologists offered their services, suggesting that perhaps it would be possible to classify untrained men by means of intelligence tests. It was proposed that the methods be given a fair trial, at least, in a selected number of military camps.

Even before the military authorities had consented to the experiment, it became apparent that, with so many thousands of men to be examined, individual examinations of the Binet type would require more time and a larger number of examiners than could be obtained for the purpose. A group of psychologists met, therefore, at the Training School, Vineland, New Jersey, to prepare methods and materials for examining recruits in large numbers. Pro-

fessor Terman, being a member of this group, explained the mechanical changes and stencils by which Otis had rendered group examinations practical, and these features were incorporated in the tests for recruits. The plan worked out by this committee involved the giving of group examinations to entire regiments and the use of individual tests upon those men whose scores in the group tests were unsatisfactory.

Psychological examining organizations were set up in four of the National Army cantonments in September, 1917. By the first of January, 1918, the military authorities had been convinced of the value of the classifications made by the psychologists, and had recommended the extension of the work to all camps.[1] The experience in the first months had demonstrated, however, that other methods must be developed for measuring the intelligence of men who could not read and write, for it had been discovered that not more than four out of five of the recruits had sufficient reading ability to deal with the group test worked out at Vineland. On the basis of the results already obtained in the army and in other smaller experiments, a new group test for illiterate men was developed, and the test

[1] The writer had the honor of being in charge of the examining unit at Camp Taylor, Kentucky, during this experimental period. The other three camps at which the methods and materials were tried out were Camp Dix, Camp Lee, and Camp Upton. For a complete history of mental testing in the Army, the reader is referred to Volume XV of the Memoirs of the National Academy of Sciences, entitled *Psychological Examining in the United States Army*, edited by Professor Robert M. Yerkes, who had charge of the organization and conduct of the entire enterprise from the very beginning to the end of the war.

for men who could read was shortened and improved. The test for illiterates was called Test Beta, while the revised test for literate men was called Test Alpha.

Test Beta was in many respects similar to the previous group tests, except that pictures and drawings were used as materials instead of printed words. The instructions were given in pantomime and by gestures rather than in spoken words. The examiner stood before the group of men to be examined and demonstrated on a specially prepared blackboard what the men were to do on their papers. On the blackboard had been painted the same type of pictures and designs as were printed on the papers in the hands of the recruits. After the examiner's demonstration, an orderly did to the designs on the board exactly what the examiner had indicated should be done. The examiner then commended the orderly for having followed instructions and signaled to the men to do the same things on their papers. The tasks to be done were of varied types, including the threading of a page of mazes, substituting appropriate symbols for Arabic numbers according to a printed key, checking the identity of paired series of numerals, completing incomplete drawings of common objects, and other simple tasks.

Group intelligence tests after the war. — The intelligence ratings supplied by the psychologists were extremely useful to those army officers who were interested in knowing how to use their men most effectively. Statistical studies showed that the psychologist's report on the intellectual strength of the men, made on the second or third day they were in camp, was more accurate than the average officer's judgment of his men after directing their work

constantly for several months. It is not surprising, therefore, that these same psychologists should turn their attention to measuring the intelligence of school children as soon as they were free from service with the Army. The military experiments had demonstrated the possibility of objectively measuring the intelligence of large groups of individuals, and educational workers were quick to see the possibilities of such tests in their field.

By a grant of $25,000 from the General Education Board, a group of psychologists was enabled to begin soon after the close of the war to modify the tests used in the army, making them suitable for use in schools. After very thorough scientific study, the National Intelligence Tests were produced. These are available in two groups, Scale A and Scale B, both of which are for the measurement of pupils in grades 4 to 8 of the public schools. Each scale is available in several different but equivalent forms. Although Scale A has thus far been used more extensively than Scale B, there is little evidence that one is in itself better than the other.

In addition to the coöperative work in developing the National Tests, many individual psychologists have prepared and published group intelligence tests for use in schools. Professor Dearborn at Harvard University and Professor Haggerty at Minnesota have each prepared one series of tests for the primary grades and another for the higher elementary grades. Both Professor Terman and Dr. W. S. Miller have published excellent test booklets for measuring the abilities of high-school students. Professor Thorndike has developed exceptionally effective intelligence examinations for measuring first-year college students.

The writer published in 1920 a group test (Mentimeters) to be used in all grades from the elementary school through the college. Many other excellent group tests are now available for testing mental abilities at various levels in the schools.

What intelligence tests should one use? — The answer to this question is, "Use that test or group of tests which will best serve the purpose without costing more than you are willing to pay." In other words, one's purposes and purse determine what mental tests one should use. Where there is need for information of maximum accuracy about the ability of each pupil in a large group, one or more group tests of intelligence should be given first, and then all doubtful cases should be carefully examined by a trained psychologist using the Stanford or the Herring Revision of the Binet Tests. It would be well to supplement these general ability tests by other tests in more specific fields of skill or knowledge, if the purpose justifies the expense. In some instances, two or three group tests will supply all the information necessary. It may even occur, in some cases, that one good group test will adequately meet all requirements. It is impossible, however, and will probably remain impossible, for one to say that any particular intelligence test is best for all purposes.

Reliability is certainly one of the cardinal virtues to be sought in choosing any test. If there is no reasonably close agreement between two trials of the same test, there is little point in using the test. The most reliable test available in one field may, however, be less dependable in another field. The fact that tests of such social virtues as trustworthiness, for example, are less reliable than the best tests

of ability in arithmetic should not excuse one for employing arithmetic tests of less than maximum reliability. One should know that the test he is employing is the most dependable instrument available for measuring the characteristic in question.

Not only should a test agree with itself, but it should also be known to measure the desired characteristic. One should use the most valid test available rather than just any test which happens to be called an intelligence test. The fact that a test is known to have a high correlation with ability to earn promotions in the military service does not prove that it would be the best test to use in determining promotion in school work. It is probable that in forecasting the child's general capacity for mathematics, one should use a different group of tests from that used in predicting his general ability to study foreign languages. The intelligence which would enable a pupil to be successful in political or social life may perhaps be measured by a different test from that used in measuring the intelligence that would enable him to be successful in academic and professional life. In any case, one should know that he is employing the test which has proved to be most valid as a measure of the quality he desires.

Still another factor should be considered in choosing an intelligence test. One should choose the test which will enable him to make the comparisons which are most desirable. Some particular test may be the best in classifying pupils of the intermediate grades but be unsatisfactory for the purposes of one who wished to compare the abilities of high-school pupils with those of primary-grade pupils. On the other hand, the test which will best

reveal the differences between elementary, high-school, and college pupils may not be available to the principal who wishes to compare the abilities of his seventh-grade pupils with published reports of seventh-grade pupils elsewhere. It would be absurd to measure pupils on scale X when one desired to make direct comparisons with pupils measured on scale Y. It would likewise be futile for one to employ a test that was applicable in the primary grades only, if his purpose was to measure the amount of change in the abilities of individual pupils as they go on into the intermediate grades and the high school. When one intends to repeat a measurement within a year, it would be wise to choose, if possible, a test having at least one alternate but equivalent form.

Summary. — Individual psychological examinations were developed in France by Binet for the purpose of detecting feeble-minded children. In America these tests were applied to hundreds of school children and were modified to give a more accurate picture of the intellectual powers of those tested. Under the stimulus of the growing popularity of objective examinations in the school subjects, group methods of intellectual examination were developed. The wholesale use of intelligence tests in the military camps gave an impetus to their development which has resulted in the publication of dozens of group intelligence tests. The number of such tests at present available makes it desirable for one to choose carefully which to use in a given case. That test or group of tests which serves one's purposes most effectively and most economically should be selected. The character of the comparisons one plans to make is important in determining one's choice of tests.

QUESTIONS AND EXERCISES

1. Can you suggest any explanation of the fact that American children are able to identify coins of various denominations at an earlier age than French children?

2. If A has an I. Q. of 80, B an I. Q. of 100, and C an I. Q. of 120, in what sense may one say that the difference between A and B is equal to the difference between B and C? Might one infer from this that the contribution to be made to the world by B will probably be just as much greater than the contribution made by A as the contribution made by C will be greater than that made by B? Why?

3. What elements contribute to make an individual psychological examination of a pupil more expensive than a group intelligence test? What advantages can be claimed for employing an individual examination rather than a group test?

4. Suppose tests A, B, and C have exactly the same coefficients of correlation with the Stanford Revision of the Binet Scale, while test A has a high correlation with test B and a low correlation with test C. Would it be more advisable to give test B or test C as a second test, following the use of test A as a first test? Why?

5. If two tests, A and B, have equal coefficients of correlation with the Stanford Revision of the Binet Scale, but A costs only three-fourths as much as B, what additional information would be required in determining whether to give A or B in your school?

REFERENCES FOR READING

E. B. Huey, *Backward and Feeble-Minded Children.*

Kite and Goddard, *The Development of Intelligence in Children.*

Pintner and Paterson, *A Scale of Performance Tests.*

L. M. Terman, *The Measurement of Intelligence.*

Yerkes, Bridges, and Hardwick, *A Point Scale for Measuring Mental Ability.*

Yerkes, et al., *Measurements of Intelligence in the United States Army.* Washington, Government Printing Office, Memoirs of the National Academy of Sciences, Vol. XV.

CHAPTER XIX

EDUCATIONAL USES OF INTELLIGENCE TESTS

ABSTRACT discussions of the nature of intelligence have been less frequent during the past few years than concrete statements regarding the amounts of it possessed by various groups of people. The extensive use of intelligence tests, frequently without any very evident consideration for the implications of the word "intelligence," has brought criticisms against many enthusiastic workers. A brilliant psychologist, in the course of a recent article in which he happened to speak of "general intelligence," added facetiously, "which may most safely be defined as that which standard intelligence tests measure." [1] Whether the tests measure intelligence or some other characteristic, that which they measure is useful to pupils and helps them to make progress through the schools. It will be helpful to the average teacher, however, to consider what the tests measure as well as what to do with the results.

What do intelligence tests measure? — It is not certain that all of the so-called "intelligence tests" actually measure intelligence. There is considerable difference of opinion as to what the essential characteristics of intelligence are. Some writers hold that it is the *present sum* of one's total

[1] K. Dunlap, "The reading of character from external signs," *The Scientific Monthly*, Vol. XV, No. 2, p. 154.

knowledge and mental ability, while others limit it to the *innate capacity* for knowledge and ability. One excellent authority defines intelligence as the "power of good responses." Another says it is ability "to carry on abstract thinking." Still others describe it as "power of the individual to adjust himself to new environment," "ability to profit by experience," or "ability to learn."[1] The present writer believes that the essential element in intelligence is the ability to see and to deal with the meanings and relationships of things. Regardless of what intelligence itself may be, the immediate problem concerns the characteristics measured by "intelligence tests."

The methods of validation used by those who have developed the group intelligence tests indicate that success in school is a good index of the ability which the tests are designed to measure. In general, each type of mental problem has found its way into the list of "intelligence tests" through some study of the closeness with which its results agree with school marks, with teachers' ratings, or with other intelligence tests. In a few instances there has been a distinct attempt on the part of a psychologist to devise a new type of test that would call for the exercise of a special capacity or ability believed by him to be indicative of intelligence. In almost every case, however, high correlation with success in school work has been advanced as one evidence of the validity of the test.

An examination of the test blanks frequently leads those who are unfamiliar with them to express surprise at the

[1] The various points of view are unusually well stated in a symposium on "Intelligence and Its Measurement," in the *Journal of Educational Psychology*, Vol. XII, pp. 123–158, 195–216, 271–275.

narrowness of the field covered. The popular understanding of intellectual ability includes a much wider group of powers than those measured by the tests. Recognition of the narrowness of these tests is growing among those who have attempted to use them. This is seen in the increasing tendency to use one group of such tests for measuring the abilities of salesmen, another group for measuring the intelligence of executive officers, and still another for measuring the mental alertness of clerical workers. Special types of intelligence are, from a practical point of view, as important as or perhaps even more important than " general intelligence."

" School intelligence tests " might be a more accurate name for the tests that show high correlations with school success. Ability to succeed in school is certainly the chief characteristic teachers are trying to predict when they apply the tests to their pupils. Furthermore, it is evident from every study of their correlations that they correspond more closely with ability in reading and language than with ability in other school subjects. The proportion of language and reading tests contained in mental test booklets varies from somewhat less than fifty per cent to fully one hundred per cent. It seems only fair to acknowledge that the tests now available are better described as " academic ability tests " than as " general intelligence tests." If such a change were made in the name of the tests, many of the questions now raised by teachers as to their uses would be answered. Whether such a change in name is made or not, one should bear in mind that the present tests measure primarily the academic phases of intelligence.

In acknowledging that available tests do not directly measure all types of intellectual ability, one does not deny that other mental powers are closely related to those measured. The writer distinguishes at least three large groups of intellectual interests: the relations between inanimate material things, the relations between people and living things, and the relations between symbols and ideas. Academic ability tests are concerned chiefly with the relations of symbols and ideas. Since most of the symbols and ideas refer to people and things and their relations with each other, there is a strong tendency for one who is highly successful in dealing with the relations of words and ideas to be reasonably successful also in dealing with the relations of people and of materials. Ability to deal with the relations of material things does not, however, so certainly imply ability to deal with the relations of their symbols. It is therefore less common for one who has well-developed mechanical intelligence to have academic interests and ability than it is for one with high academic intelligence to have fair mechanical and social interests. Figure 44 illustrates the writer's conception of the relationships between general intelligence and the abilities measured by intelligence tests.

Intelligence tests in educational guidance. — Assuming that the tests now available are measures of general "academic ability," one may point out the value of such tests in helping pupils to avoid academic disaster. The experience of the psychologists in the army seemed to indicate that almost one-fourth of the adult men of the United States were not able to use reading and writing effectively as tools for communicating thoughts. Study of the

recruits in individual examinations convinced the writer that it will be almost hopeless to try to reduce the proportion of illiteracy below fifteen or sixteen per cent. According

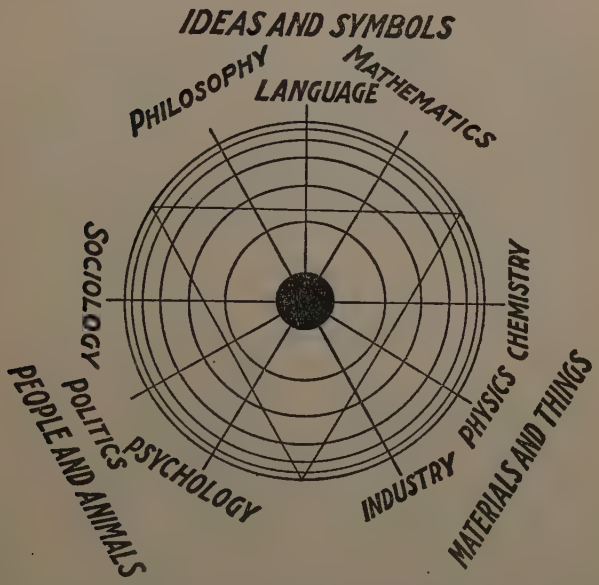


FIG. 44. — SCHEMATIC REPRESENTATION OF A VARIETY OF INTELLECTUAL INTERESTS

The so-called "intelligence tests" do not measure all phases of intelligence equally well. The upper side of the triangle, academic interests, receives an undue amount of attention in present tests. The person who sees and employs the meanings and relations of inanimate things is intelligent, and so is also the person who understands and employs the relations and motives of people; but we have tended to credit with greatest intelligence those who work with the significations and relations of words or symbols.

The above diagram is inadequate in that many lines of interests are not represented. General intelligence may perhaps be represented more accurately by a sphere than by a circle or triangle. It certainly cannot be restricted to the narrow academic field measured by the group intelligence tests.

The concentric rings suggest that an equal amount (area) of *general* growth during a given period would show less and less growth in any one direction. As one's interests become more numerous and more well-defined, one's abilities in directions not related to one's specialty almost stop developing.

to this estimate, approximately one-sixth of the population of our country will never be able to use written and printed English as an effective tool in civic and social life. The intelligence tests serve admirably as one means of detecting these persons early in their school life, in order that other tools may be developed for them by the school.

The scheme of education that has grown up in America suggests that school training is primarily for those who are planning to enter the "learned professions." The elementary school prepares for high school, and the high school prepares for college. Those who have no interest in this line of progress show signs of practical intelligence when they drop out of school and go to work as soon as the law will allow them to do so. Almost every state constitution declares that the schools shall offer equal opportunity to all children in the state, and yet in most states the academic type of mind is thus far the only type adequately provided for at public expense. Those pupils whose intellectual interests are in other lines of development find few avenues of progress open to them. As a result, thousands of pupils carry home at the end of each term reports which say they are "FAILURES." Now that reasonably valid, objective tests are available for discovering pupils who have little academic ability, the school and the state are guilty of a great wrong to each of these children as long as they fail to provide them with school work in which they can be "SUCCESSES."

For those pupils whose academic abilities enable them to enter successfully the high schools and colleges, intellectual test scores are valuable as a basis for advice regarding elective studies. It is not difficult to determine by means

of these tests the minimum degree of mental alertness required for success in each study, and then to base one's advice to students on definite knowledge of their abilities and of the requirements of the study. Certain of the more progressive colleges are beginning to recognize that mental ability is more significant of academic success than is memory of work previously done in preparatory school. An increasing number of students is therefore admitted to college each year on the basis of fitness to do college work, as certified by their scores on intelligence tests. The fact that intelligence tests are so largely academic in character renders them all the more suitable for use in such situations.

The discovery of the relative constancy of Intelligence Quotients led many persons to hope that the intelligence tests would become a primary means of determining future vocations for young people. The thought was that if one could only discover the I. Q. necessary for success in each vocation, it would be possible to measure a pupil at any stage of his development and to tell at once the vocation into which he should go. Experiments have demonstrated, however, that success in a given vocation does not seem to be confined to those of any particular I. Q. It is true that successful lawyers, physicians, and accountants usually have more than a moderate amount of academic intellectual ability. There is also a tendency for each vocation to have associated with it a minimum level of intelligence, below which success is not usual. But if a child has more than the minimum intelligence for success in a given vocation, it is by no means certain that he might not be even more successful if he entered some calling with a still lower intellectual minimum. In other words, vocational guidance

in general must be based more largely on tests of specific vocational interests than on tests of general academic capacity.

Intelligence tests in educational organization. — Ideally, every child should move forward in his education just as rapidly as his abilities will permit. Individual attention to each child in each subject is the ideal toward which school organization is directed. But individual assignments and individual recitations would be too expensive for the public schools and would also remove the social stimulus which comes from class recitations. The classification of pupils in public schools therefore resolves itself into a problem of grouping in the same classes those pupils who have the same abilities and who are moving forward at the same rate. Intelligence tests help to reveal those pupils who have the same general academic abilities, and to indicate those who can move forward at about the same rate, while the standard educational tests measure the degree of progress each pupil has made in the specific school subjects.

Special classes for slow-progress pupils, who are three or more years older than the normal ages for their grades, have been provided in progressive school systems for many years, and are now required by law in many states. In recent years an increasing number of schools have also provided special classes for rapid-progress pupils whose abilities place them in grades higher than those in which pupils of their ages would normally be found. In such special classes, not more than fifteen pupils are enrolled, and each pupil receives individual assignments from the teacher. The removal of the extremely rapid and the

extremely slow has been found to be beneficial to the average pupils who are left in the original classes, as well as to the pupils who are removed and placed in the classes where they may receive individual attention. The net result of such homogeneous grouping is that all pupils make more rapid progress and do more thorough work than under the old system.

One objection to the establishment of such classes is that some children may be embarrassed by being assigned to a " special " class. In actual practice this feature of the problem has not proved serious. It would be wise, however, to secure if possible the same general improvement in the total situation without resorting to special groups. If the grouping together of pupils having the same abilities can be obtained by the normal working of the classification machinery and without actually making special cases of a few pupils, both pupils and teachers will have a more wholesome attitude toward their work. There is no need to designate certain pupils as being " unusual " just because they happen to be at the extreme ends of a distribution for ability. It is just as " normal " for a distribution to have small numbers of pupils at the extremes as to have large numbers near the average.

The more serious objection has frequently been raised that any grouping together of pupils of like ability will represent an artificial situation unlike life outside the school, where all types and degrees of ability are to be found in practically every group. It is claimed that uniformity of ability in a class would tend to destroy the social stimulus of competition and the opportunity to develop leadership. Perhaps it is unnecessary to point out that

this objection can be raised against all attempts at grading pupils, and that it only applies more strongly to the plan here outlined because the grouping by objective tests is more accurate than the usual grouping by teachers' examinations. The answer to the objection is found in the fact that the greater the similarity of ability in a group, the greater is the opportunity for each pupil to stand sometimes at the head of the class and to be the leader. It is much more stimulating to work in competition with one's equals than to compete with those whom one can never overtake. Furthermore, absolute uniformity of ability cannot be obtained in any group. There will always be ample opportunity for the development of leadership and for the social stimulus of competition within any homogeneous group in which a pupil may find himself.

It has been suggested by some superintendents and principals that pupils should be classified in school according to their scores in intelligence tests. Such a plan is entirely too simple for practical use, except in admitting pupils to the first school grade, and even there it may well be modified by other considerations. When pupils have made some progress in the various school subjects, the amount of their achievement should certainly be considered in grouping them for further progress. *Pupils should be placed in a given class because of their readiness to do the work of that class.* Readiness, in terms of subject matter already mastered, may be determined most accurately from scores in standard educational tests. Readiness, in terms of mental power to accomplish the work to be done, may be most certainly determined by scores in the intelligence tests. Readiness, in terms of the speed at which the work is to

be done, may best be determined from intelligence quotients. Neglect of any one of these three indices of readiness will result in less satisfactory classifications than are now possible.

The question should probably be raised at this point as to whether the grade or class distinctions used in the public schools are really essential. Might it not be worth while to " promote " every child each year (if not more frequently), in the sense that his grouping for school work be studied annually and if possible improved? Instead of pretending that all elementary school graduates have the same amounts of ability, would it not be better to issue, at the end of the sixth grade, for example, certificates of school attendance, showing just how well each pupil can perform various types of school tasks, in terms of definite objective scales? Such a certificate, or an Educational Achievement Chart of the type represented on the following page, would be infinitely more significant than an ordinary elementary school diploma now is. The child whose intellectual interests were chiefly concerned with tools and mechanical devices should also receive a certificate at the end of the six years. In his case there would appear definite records of his abilities in those lines along which the schools had found it possible to train him.

It is perhaps too much to expect the abolition of school grades and the " failure-to-be-promoted " system, at least in the immediate future. One may confidently expect, however, the general adoption at an early date of objective bases for classifying pupils in more homogeneous groups. As long as one is attempting to group together pupils of similar abilities, one should seek to employ the most effective ways to accomplish his purpose.

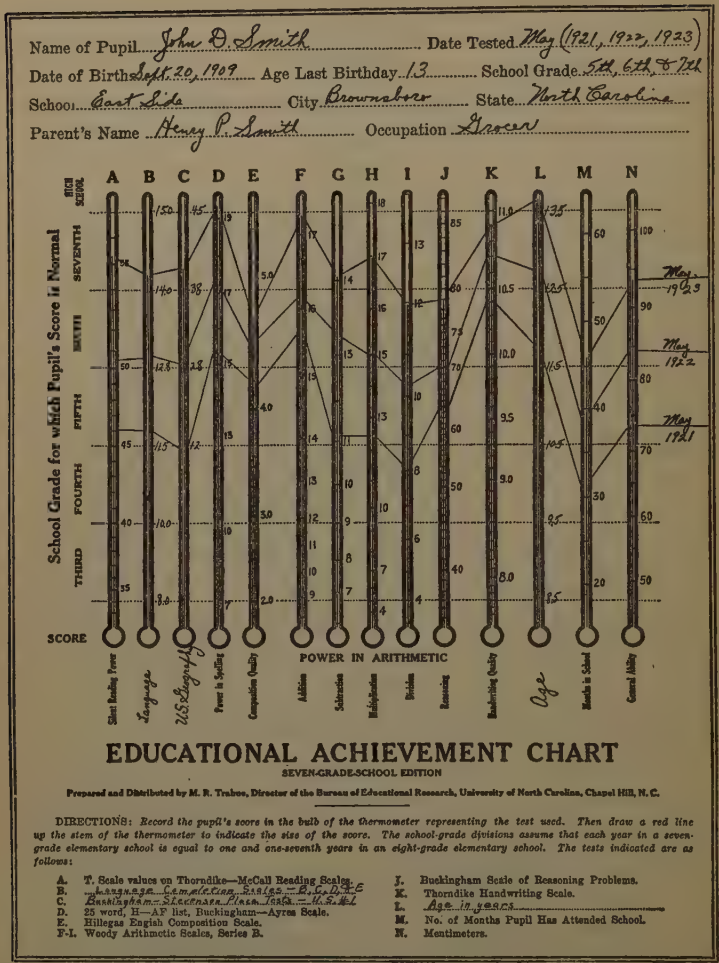

Name of Pupil John D. Smith Date Tested May (1921, 1922, 1923)
Date of Birth Sept. 20, 1909 Age Last Birthday 13 School Grade 5th, 6th & 7th
School East Side City Brownsboro State North Carolina
Parent's Name Henry P. Smith Occupation Grocer
School Grade for which Pupil's Score is Normal
HIGH SCHOOL
SEVENTH
SIXTH
FIFTH
FOURTH
THIRD
A B C D E F G H I J K L M N
May 1923
May 1922
May 1921
SCORE
POWER IN ARITHMETIC
Silent Reading Power
Language
U.S. Geography
Power in Spelling
Composition Quality
Addition
Subtraction
Multiplication
Division
Reasoning
Handwriting Quality
Age
Months in School
General Ability
EDUCATIONAL ACHIEVEMENT CHART
SEVEN-GRADE-SCHOOL EDITION
Prepared and Distributed by M. R. Trabue, Director of the Bureau of Educational Research, University of North Carolina, Chapel Hill, N. C.
DIRECTIONS: Record the pupil's score in the bulb of the thermometer representing the test used. Then draw a red line up the stem of the thermometer to indicate the size of the score. The school-grade divisions assume that each year in a seven-grade elementary school is equal to one and one-seventh years in an eight-grade elementary school. The tests indicated are as follows:
A. T. Scale values on Thorndike—McCall Reading Scales.
B. Language Completion Scales — B, C, D, E
C. Buckingham—Stockmayer Place Tests — U.S. #1
D. 25 word, H—AF list, Buckingham—Ayres Scale.
E. Hillegas English Composition Scale.
F-I. Woody Arithmetic Scales, Series B.
J. Buckingham Scale of Reasoning Problems.
K. Thorndike Handwriting Scale.
L. Age in years
M. No. of Months Pupil Has Attended School.
N. Mentimeters.

FIG. 45. — A graphic chart showing educational test scores. Such a chart may be used to show the scores of an individual pupil or the average scores of an entire class. The same chart may be used year after year to show progress by drawing connecting lines horizontally, as illustrated.

Combinations of bases for classifying pupils. — In small elementary schools having not more than eight or ten rooms, the problem of selecting a basis for the classification of pupils is much simpler than in large schools. The small school is limited in the number of classes possible and therefore cannot hope to obtain as homogeneous groupings as are possible in the large school. Even in the small school, having not more than one room for each grade, it is nevertheless desirable to divide each grade into at least two recitation groups on the basis of scores in educational and intelligence tests. Although this combination of objective test scores is more reliable in general than the judgments of teachers, the time has not yet arrived when the teacher may be relieved of making judgments regarding such matters. The results of the educational and intelligence tests should always be modified in some degree by teachers' estimates of each pupil's readiness for the work to be done.

In large schools, the greater number of classes or sections in each grade increases the opportunity for securing in each recitation group pupils who are very nearly equal in their abilities. Some attempts have been made to group together pupils having high intelligence quotients, regardless of their mental and chronological ages. Where pupils have been grouped for class work on the basis of their intelligence quotients alone, the scheme has failed on account of the wide differences in mental abilities and educational advancement of pupils having the same I. Q. Two pupils may each have an I. Q. of 125, but one may be only six years old and just entering the first grade, while the other may be twelve years old and ready to graduate from

the elementary school. In a group having the same educational and intellectual scores, however, those who are youngest have the highest I. Q.'s and will make the most rapid progress, while those who are oldest will make the slowest progress. The intelligence quotients are therefore extremely useful in subdividing groups having the same present abilities, but intelligence quotients cannot be used independently as a basis of classification.

In order to avoid making many " special " classes, which call unnecessary attention to the extremely bright and to the extremely dull pupils, and in order to obtain for all pupils the benefits of classification by mental-age scores and by intelligence quotients, without emphasizing them, the writer has for several years advocated the following combination of bases for grouping pupils homogeneously. Let admission into a given grade be on the basis of a combination, including scores on standard educational tests, scores on intelligence tests, and the judgments of teachers as to the pupil's readiness for the work of the grade. In other words, let promotion from one academic grade to the next be based on all the most objective evidences of readiness obtainable. Having pupils of the same general abilities and educational achievements in a given grade, let recitation sections within the grade be formed by grouping together those who have the same degree of physical maturity or chronological age.

A school classified on this combination of bases would show in each recitation section pupils of approximately the same degree of advancement in subject matter, of about the same mental maturity, and of similar powers as regards the rapidity with which they could take school training.

Mental ages, educational achievements, and intelligence quotients would be approximately uniform within each class although there would be nothing to call the special attention, even of the teachers, to the intelligence quotients. A parent who inquired why his child was placed in a given class could be told plainly and truthfully that the child was placed in the grade to which all children of equal attainments belonged, and in the recitation section to which other children of the same physical maturity were assigned. The parent could have pointed out for him the advantages of having his child in a class with other children having the same social interests and physical development, even though the chief advantage to the school would be that the pupils in each recitation section would have similar rates of progress.

School authorities should whenever possible avoid discussing intelligence quotients with the public. They are the professional tools of the school authorities and should no more be made public property than should the blood pressure readings obtained by a physician. There is probably no fact about a child that so clearly indicates his probable future success as does his intelligence quotient. Those who are held responsible by the state for the guidance and training of children cannot afford to be ignorant of the I. Q. of any child in their care; but it is usually unwise to arouse the emotions of parents by telling them the degree of intellectual capacity their children have inherited. Whenever it becomes more widely recognized that intelligence, or academic ability, is not a moral quality for which one may personally be held responsible, it may perhaps become possible for persons to discuss their mental characteristics without emotion.

Proper classification is a beginning, not an end. — When pupils have been properly classified on the basis of mental ages, physiological maturity, vocational interests, and other considerations, teachers and supervisors have before them an opportunity, not an achievement. Unless one undertakes seriously to work out appropriate adjustments of the content and method of the course of study to the abilities and interests of the group so carefully formed, the time and effort required to make the classification have been wasted. The average teacher will probably not be able to make the needed adjustments without careful supervision. It is so much easier to give a rapid-progress group " more of the same " than it is to work out new and appropriate materials. Only by constant attention and effort may one make the nice adjustments between ability and assignments which are the chief objectives of homogeneous grouping.

One of the greatest obstacles to the proper adjustment of school work and ability has been the tendency for teachers to consider those school subjects which lead toward academic life as being the most worthy and respectable. It may be true that ability to deal with the relations of symbols usually involves a somewhat wider field of interests than is involved by such abilities as make up mechanical intelligence, but the one type of ability is certainly no more respectable than the other. It is just as worthy for one child to study market-gardening as it is for another to study the classics, provided each is working at the thing in which he has real interest and ability. On the other hand, the school itself is inadequate and unworthy if it says to the child with the market-gardening interest, " You may

not have a high-school education except by studying the classics."

The accomplishment ratio. — Success in arithmetic or in reading is not dependent on mental ability only. It is probable that people differ as decidedly in their inclinations to work as in their abilities. It is not certain, therefore, that a pupil with a high intelligence score will actually accomplish more than a pupil with a lower intelligence score. The bright pupil may actually accomplish less, considering their relative abilities, than is accomplished by the dull pupil. In order to obtain an estimate of how well each pupil is living up to his own mental possibilities, efforts have recently been made to obtain Accomplishment Quotients or Accomplishment Ratios.[1]

Those who seek accomplishment ratios find it necessary first of all to turn the scores for educational tests into age indices, showing for what age each child's score is typical. From the score of a spelling test, therefore, one obtains a Spelling Age, and from a reading test, a Reading Age. The average of all these subject indices is called an Educational Age, and is directly comparable to the Mental Age obtained from the intelligence tests. The ratio between the Educational Age and the Mental Age is called the Accomplishment Ratio. It indicates how nearly the pupil is living up educationally to his intellectual possibilities. It then becomes the task of the teacher to bring the child's Accomplishment Ratio (A. R.) as nearly as possible to 1.00, — that is, to make his Educational Age as nearly equal as possible to his Mental Age.

[1] See especially R. Franzen, "The Accomplishment Quotient," *Teachers College Record.*

One of the important uses for these derived quotients or ratios is to determine for each pupil and in each subject the point at which it is wise to stop trying to secure greater achievements. As a pupil's Reading Ratio approaches unity (1.00), as indicated by the fact that his Reading Age approaches his Mental Age, it is probable that the "law of diminishing returns" begins to act on the instruction he receives. If this is the case, then the teacher would do well to spend much of her effort helping a child in that subject in which his Educational Age index is farthest below his Mental Age. In order to raise his Arithmetic Age, his Spelling Age, or any other Subject Age to a point higher than his Mental Age, a child would usually have to put forth extraordinary efforts.

The helpfulness of these ratios as guides to teachers becomes clear when it is known that those pupils who have the highest intelligence quotients usually have the lowest accomplishment ratios. This means that teachers have not secured as adequate results from their bright pupils in proportion to their abilities, as they have from their slower pupils, probably because they have not recognized the true limits of these abilities in either case. Failure to recognize high academic ability and to keep it supplied with work of a stimulating nature, results in an enormous amount of educational waste. The ratio method of relating educational achievements to general academic ability, promises to help materially in calling attention to and in conserving the abilities of the intellectual leaders of the future. One should not forget, however, that the scores from the so-called "intelligence tests" are also greatly influenced by the environment and training the pupil has had.

Summary. — Intelligence tests probably measure the academic phases of intelligence much more accurately than they measure the social and mechanical abilities. The intelligence tests are especially valuable in educational work because they do measure educational readiness rather than general ability. Educational tests show the amount of specific knowledge and special ability, while intelligence tests show general academic power and the speed at which progress can be made. Both types of tests, teachers' judgments, and maturity of pupils should be used in making up homogeneous groups of pupils. Grouping pupils of like abilities and outlook is not in itself a virtue, except as adjustments of curriculum and method are made to meet the new opportunities thus created for more adequate instruction. The relation of a pupil's educational achievements to his intellectual powers serves as a practical basis for directing teachers and supervisors in their efforts to develop the abilities of each pupil.

QUESTIONS AND EXERCISES

1. What difference in point of view do you see between the definition of intelligence as "power to carry on abstract thinking," and its definition as "power of the individual to adjust himself to new environment"?

2. What definition of intelligence does the maker of an intelligence test adopt when he checks the validity of his test by comparing its results with the progress pupils have made in school?

3. Would ability to remember accurately what is contained in a show-window or a storeroom prove the presence of great intellectual capacity?

4. May one be intelligent about an automobile without being intelligent about school organization? Does being "intelligent" about a thing mean the same as being informed about it?

5. What do you find in your knowledge of the history of secondary education which would explain the tendency to build up academic courses and to neglect vocational courses in the public high schools?

6. What objections might be raised by pupils themselves to being classified in "Special" classes?

7. Would it be wisest, if one single basis must be used, to group pupils by mental ages, by intelligence quotients, by educational ages, or by chronological ages?

8. Why should a pupil who works earnestly and does his best receive a mark of "failure" in school work?

9. Is a teacher appointed because of her skill in classifying pupils, or for some other reason? Who should be held primarily responsible for the proper classification of pupils?

10. If it were discovered that a 12-year-old pupil had the following "age scores," to what school subjects should the teacher give her attention in order to aid the boy as much as possible?

Mental age	12 years
Reading age	14 years
Spelling age	12 years
Addition age	11 years
Subtraction age	10 years
Multiplication age	11 years
Division age	10 years
Arithmetic Reasoning age	13 years
Writing age	12 years
Composition age	13 years
Language age	14 years

REFERENCES FOR READING

W. F. Book, *The Intelligence of High School Seniors.*
V. E. Dickson, *Mental Tests and the Classroom Teacher.*
H. C. Hines, *Measuring Intelligence.*
H. H. Goddard, *Human Efficiency and Levels of Intelligence.*
L. M. Terman, *The Intelligence of School Children.*
Terman, et al., *Intelligence Tests and School Reorganization.*
Thorndike, et al., *Intelligence Tests and Their Use.*
Ben D. Wood, *Measurement in Higher Education.*

CHAPTER XX

THE DERIVATION OF EDUCATIONAL SCALES

WHOLESALE use of a scale is not sufficient evidence to prove that it is built on scientific principles or that it measures accurately the quality it was intended to measure. In altogether too many cases, teachers and school officials have administered a given test to their pupils for no other reason than that the same test had been used in another school with which they were acquainted. Although there is a distinct advantage in having records from other schools with which to compare, it is not advisable to encourage and extend the use of a poorly constructed test when more accurate and helpful tests are available in the same field.

This chapter is not intended as a guide to those who wish to prepare and standardize new educational scales. Its only purpose is to show in a general way the methods that have been used in the derivation of some of the best scales now available. When he has reached an understanding of the methods involved, the reader should be able to examine a test intelligently and to decide for himself whether any given educational scale is constructed on sound statistical principles. A brief discussion cannot include all of the details and refinements that are being developed constantly by the workers in this field, but it will serve to

acquaint the reader with the general procedure employed, and it may enable him to avoid wasting much time and effort in the giving of unscientific tests.

The desirability of equal steps on a scale. — It is not necessary that a scale have equal steps in order that it may be useful. In the absence of a yardstick with equal units (inches) marked off on it, one might still use with reasonable accuracy a stick on which there were marks at $1\frac{1}{2}$ inches, 2.6 inches, $4\frac{3}{4}$ inches, 5.7 inches, 7.2 inches, and so on. It is much more convenient, however, to use a stick having one mark at 1 inch, another at 2 inches, a third at 3 inches, and so on at regular intervals. Irregular intervals or steps do not prevent accurate measurements, but they make such measurements more difficult to obtain.

The advantages of equal intervals or steps in an educational measuring scale are even greater than in a scale for such physical qualities as weight and length. People are relatively familiar with physical measurements and are less likely to misinterpret the scores than they are when such an unfamiliar quality as reading ability is being measured. Almost any school pupil would be able to find and state correctly the weight of a 16-pound sack of flour by the familiar process of balancing it with one after another of a series of graduated weights, even though these were at such irregular weight-intervals as 11, 13, 14, 14, 15, 16, 16, 17, 19, and 36 pounds. When the above numerals refer to the values of questions in reading, however, there are excellent teachers who do not detect the absurdity of giving a score of 99 to the pupil who masters all the first seven selections ($11 + 13 + 14 + 14 + 15 + 16 + 16 = 99$),

and a score of 83 to the pupil who masters only the first six. Even when the intervals between consecutive values in a scale are equal, the average teacher has more difficulty in finding and tabulating the scores from educational scales than she has in interpreting scores from physical measurements. Because of the added difficulties which an irregular scale always introduces, it is desirable to use educational scales with as nearly uniform intervals between the values of the different scale elements as it is possible to obtain.

Although equality of steps is a desirable characteristic in educational scales, it is sometimes almost impossible to obtain. The nearest approach to equality that Hillegas was able to obtain in his scale for measuring the merit of English composition was: 0, 1.83, 2.60, 3.69, 4.74, 5.85, 6.75, 7.72, 8.38, and 9.37. If no use had been made of the Hillegas scale on the ground that the steps were irregular, Hudelson would never have been able to prepare his supplement to this scale, in which the steps are quite regular: 2.07, 2.50, 3.00, 3.56, 4.03, 4.51, 5.06, 5.50, and so on. The values of scale elements are determined by statistical study of what people do when actually confronted with them, and the only way to obtain an element that will have the desired statistical value is to keep on trying out new elements until one is found that has the correct value. This " cut and try " method sometimes requires the evaluation of hundreds of items in order to find fifteen or twenty having the values one desires to include in a scale. It will be seen from this that scale building is a tedious process, requiring a large amount of patience and persistence, and a high regard for scientific accuracy.

A permanent basis desirable for measuring relative quality or difficulty. — How much more difficult is one problem in arithmetic than another? What are the relative difficulties of five spelling words? In order to deserve recognition as scientifically useful statements, the answers to such questions must be in terms of some permanent objective fact. It must be possible for competent persons, fifty or a hundred years hence, to verify the accuracy of the amounts of difference.

If one were to present the accompanying addition examples to a hundred elementary school teachers and ask them to tell how much more difficult the second is than the first, probably half of them would reply in general terms and the other half in the language of percentages. One teacher would guess that the second was "very much harder," and another that it was "twenty-five per cent harder." Perhaps neither of them would know exactly the meaning of the reply. It is doubtful whether any mathematician a hundred years from now would be able to identify accurately the amount of difference a teacher at the present time tried to indicate by saying "fifty per cent more difficult."

Add

(*a*)	(*b*)
2	17
3	2

The selection of permanent bases or units for the measurement of mental differences is even more difficult than the selection of permanent bases for measuring physical facts. The Centigrade Scale for measuring temperature, for example, employs as its unit one one-hundredth of the difference between the boiling point and the freezing point of water. This difference remains relatively constant from

year to year, and any thermometer may be verified by comparison with it.[1] The amount of difference between the degrees of addition ability possessed by two pupils cannot be used in a corresponding manner, however, for both pupils will certainly change in ability as they grow older, and the amount of change in one will probably not be equal to the amount of change in the other. Even if their rates of change were to remain the same, both pupils would in a relatively few years leave the world without the means of verifying the addition scale.

It is likewise impossible to use the difference between the average abilities of any two school grades as a unit, for present fifth-grade ability may be equal to fourth-grade ability within ten years. Nor is it possible to use the difference between the average eight-year-old and the average ten-year-old child as a unit, for the amount of training in addition given to children of these ages may in twenty years be quite different from the amount such pupils are now given. Group *averages* cannot be used to indicate the limits of a scale unit, because it is impossible, without actual measurements on an objective scale, to know that any two groups made up at different times and in different places are equal in ability.

It has been suggested that in the preparation of an addition scale, for example, a large group of pupils be tested with a considerable number of problems; that an accurate record be kept of the percentage of the whole number of children solving each problem correctly; and that one degree or unit be such a difference in difficulty as would make

[1] The reader may wish to read again at this point the discussion of the Standard English Yardstick given in Chapter I.

a ten per cent difference in the percentages of correct answers. Under such a scheme, a problem solved by forty per cent of the group would be one unit more difficult than a problem solved by fifty per cent, and at the same time one unit less difficult than a problem solved by only thirty per cent of the group.

Although this plan for measuring the differences between the difficulties of problems or of questions seems attractive, it does not work out well in actual practice. A difference in difficulty between two problems which would cause ten per cent difference in success with one group will frequently be found to cause twenty or thirty per cent difference with another group of pupils having a higher or lower median ability. Figure 46 shows how the differences obtained from the

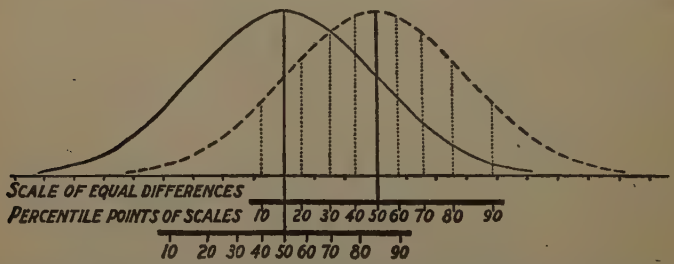


FIG. 46. — Showing inequality of differences between decile points. Note that the difference between the 40 percentile and the 60 percentile in the upper distribution is little more than the difference between the 80 percentile and the 90 percentile in the lower distribution.

study of one group of pupils would fail to hold good for another group having a higher median ability, if both groups were distributed about their medians according to the law of normal probability.

It is clear from a study of Figure 46 that the difference between a problem solved by ten per cent and a problem

solved by twenty per cent of a group, is greater than the difference between two other problems, one of which is solved by twenty per cent and the other by thirty per cent. The nearer the ten-per-cent difference between problems is to the median or fifty-percentile of a group, the smaller the interval. The difference between a ten-percentile problem and a twenty-percentile problem would be equal to the difference between a fifty-percentile problem and a sixty-percentile problem only in case the distribution of ability in the group of pupils studied could be properly represented by a rectangle. All the evidence available seems to show that the normal surface of frequency represents the distribution of ability in a grade or in an age-group much more accurately than it would be represented by a rectangle or any other type of figure.

Dispersion of a group as a measure of differences. — The most satisfactory means yet found for determining differences in the difficulty of test questions is similar to the plan just discussed, except that the amounts of difference are expressed in terms of distances along the base-line of a normal surface of frequency rather than in percentages of correct answers. The base-line of a normal surface of frequency is thought of as a linear scale along which the test problems are distributed in order according to their difficulty. The location of each question is such that if a perpendicular were erected at that point on the base-line, the surface of frequency would be cut in the same proportions as the group was divided between success and failure on that question. A problem solved by exactly half of a group would be located at the median or mid-point of the base of the surface. A problem solved by more

than half would be located below (at the left of) the median point, and a question answered by less than half the group would be located above the median, at a distance depending on the percentage of the group that had been unsuccessful.

Distances along this base-line are commonly expressed in terms of some measure of the dispersion of the group. Any point is identified as at some fraction of a Standard Deviation above or below the median, or as at so many P. E. from the median. There is no reason why the A. D. and the Q., or any other measure of dispersion, should not be used in exactly the same manner, but the S. D. and Median Deviation (P. E.) have come to be used most frequently as measures of distances from the median. Because of the frequency with which it is desirable to convert percentages of correct answers into locations on the base-line of a normal surface of frequency, Table LXIII has been prepared especially for that purpose.

Converting percentages of success into P. E. locations. — The use of Table LXIII may be illustrated by considering again the two addition examples listed on page 452. In an actual trial with 489 second-grade pupils, example *a* was solved correctly by 88.6%, and example *b* by 66.1% of the pupils. From the table, it may be seen that 88.6% of success would locate problem *a* at − 1.79 P. E. from the median, and that 66.1% would locate problem *b* at − .62 from the median. The difference in difficulty between problems *a* and *b*, as indicated by the second-grade distribution, is therefore 1.17 P. E. (1.79 − .62 = 1.17).

In the third grade, problem *a* was solved by 94.6%, which would place it 2.38 P. E. below the third-grade median; and problem *b* was solved by 81.5%, which would

TABLE LXIII

P. E. Distances from the Median Corresponding to Each Per Cent of the Normal Surface of Frequency

%	P. E. above M	%	P. E. above	%	P. E. below M	%	P. E. below M
.5	3.82	25.5	.98	50.5	.02	75.5	1.02
1.0	3.45	26.0	.95	51.0	.04	76.0	1.05
1.5	3.22	26.5	.93	51.5	.06	76.5	1.07
2.0	3.04	27.0	.91	52.0	.07	77.0	1.10
2.5	2.90	27.5	.89	52.5	.09	77.5	1.12
3.0	2.80	28.0	.86	53.0	.11	78.0	1.14
3.5	2.69	28.5	.84	53.5	.13	78.5	1.17
4.0	2.60	29.0	.82	54.0	.15	79.0	1.20
4.5	2.51	29.5	.80	54.5	.17	79.5	1.22
5.0	2.44	30.0	.78	55.0	.19	80.0	1.25
5.5	2.37	30.5	.76	55.5	.20	80.5	1.27
6.0	2.30	31.0	.73	56.0	.22	81.0	1.30
6.5	2.24	31.5	.71	56.5	.24	81.5	1.33
7.0	2.19	32.0	.69	57.0	.26	82.0	1.36
7.5	2.13	32.5	.67	57.5	.28	82.5	1.39
8.0	2.08	33.0	.65	58.0	.30	83.0	1.41
8.5	2.03	33.5	.63	58.5	.32	83.5	1.44
9.0	1.99	34.0	.61	59.0	.34	84.0	1.47
9.5	1.94	34.5	.59	59.5	.36	84.5	1.51
10.0	1.90	35.0	.57	60.0	.38	85.0	1.54
10.5	1.86	35.5	.55	60.5	.39	85.5	1.57
11.0	1.82	36.0	.53	61.0	.41	86.0	1.60
11.5	1.78	36.5	.51	61.5	.43	86.5	1.64
12.0	1.74	37.0	.49	62.0	.45	87.0	1.67
12.5	1.71	37.5	.47	62.5	.47	87.5	1.71
13.0	1.67	38.0	.45	63.0	.49	88.0	1.74
13.5	1.64	38.5	.43	63.5	.51	88.5	1.78
14.0	1.60	39.0	.41	64.0	.53	89.0	1.82
14.5	1.57	39.5	.39	64.5	.55	89.5	1.86
15.0	1.54	40.0	.38	65.0	.57	90.0	1.90
15.5	1.51	40.5	.36	65.5	.59	90.5	1.94
16.0	1.47	41.0	.34	66.0	.61	91.0	1.99
16.5	1.44	41.5	.32	66.5	.63	91.5	2.03
17.0	1.41	42.0	.30	67.0	.65	92.0	2.08
17.5	1.39	42.5	.28	67.5	.67	92.5	2.13
18.0	1.36	43.0	.26	68.0	.69	93.0	2.19
18.5	1.33	43.5	.24	68.5	.71	93.5	2.24
19.0	1.30	44.0	.22	69.0	.73	94.0	2.30
19.5	1.27	44.5	.20	69.5	.76	94.5	2.37
20.0	1.25	45.0	.19	70.0	.78	95.0	2.44
20.5	1.22	45.5	.17	70.5	.80	95.5	2.51
21.0	1.20	46.0	.15	71.0	.82	96.0	2.60
21.5	1.17	46.5	.13	71.5	.84	96.5	2.69
22.0	1.14	47.0	.11	72.0	.86	97.0	2.80
22.5	1.12	47.5	.09	72.5	.89	97.5	2.90
23.0	1.10	48.0	.07	73.0	.91	98.0	3.04
23.5	1.07	48.5	.06	73.5	.93	98.5	3.22
24.0	1.05	49.0	.04	74.0	.95	99.0	3.45
24.5	1.02	49.5	.02	74.5	.98	99.5	3.82
25.0	1.00	50.0	Median	75.0	1.00	100.0	—

indicate that its "difficulty location" is 1.33 P. E. below the median. The difference between these problems, when measured by the third-grade distribution, is therefore 1.05 P. E., which is not far from the 1.17 P. E. difference previously found by use of the second-grade distribution.

The values found for the two addition examples in each of the two distributions assist one in estimating the difference between the medians of the two grades. Problem *a* had a difficulty of −2.38 in the third grade, and of −1.79 in the second grade, so that one may subtract the second-grade value from the third-grade value and obtain .59 P. E. as the distance which separates the two medians. Problem *b* had a value of −1.33 P. E. in the third grade, and −.62 P. E. in the second, which would indicate that the difference between median second-grade ability and median third-grade ability is .61 P. E. Measurements of this difference between the second and third grades, by means of a large number of more difficult addition problems, indicates that .59 P. E. and .61 P. E. are relatively small; but the average of these two, .60 P. E., might be used if other and more complete evidence were not at hand.

Figure 47 shows the locations of both addition problems in each of the two grades, which are so adjusted as to show .60 P. E. between their medians. The reader should note that both problems are below (to the left of) the median ability in each grade, indicating that more than half of the pupils were successful.

Another method of employing Table LXIII to determine the difference between the medians of two grades or groups involves the distribution of the scores of each grade on a long list of questions or problems. In testing more than

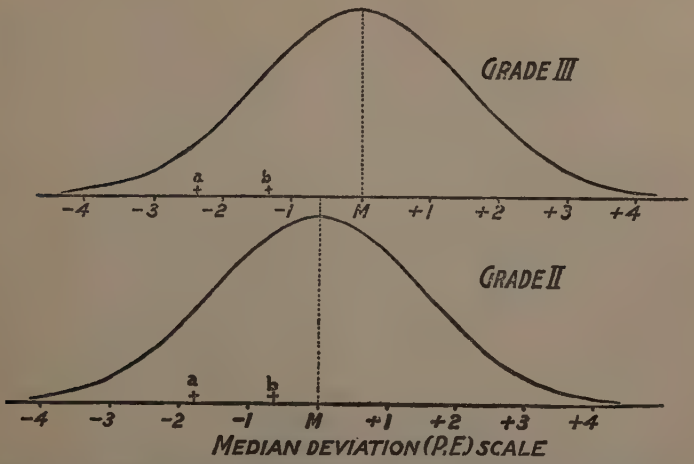


FIG. 47. — Difficulty locations of two addition problems in Grades II and III. Problem *a* was solved by 89% of the second grade and by 95% of the third grade. Problem *b* was solved by 66% of the second grade and by 81% of the third grade.

six hundred pupils in each of grades four and five, with a graded list of thirty-eight addition examples, Dr. Woody found that the median score of the fourth grade was 18.32, and the median score of the fifth grade was 23.07 problems right.[1] By calculation it was found that 91.2% of the fifth grade had a higher score than the median of the fourth grade. Table LXIII shows that this percentage would place the fourth-grade median at − 2.01 P. E. from the fifth-grade median. It was also found that 9.6% of the fourth-grade distribution was higher than the fifth-grade median, which would place the fifth-grade median (according to Table LXIII) at 1.95 P. E. above the fourth-grade

[1] C. Woody, *Measurements of Some Achievements in Arithmetic*, Teachers College, Columbia University, Contributions to Education, No. 80, p. 27.

median. Either of these two determinations, or their average, 1.98 P. E., might be taken as a measure of the difference between the medians of the two grades. Figure 48 shows by its shaded area the percentage of the fourth grade having scores higher than the median for the fifth grade.

The unit of measurement, on a scale of this type, is a statistical measure of the dispersion of a group of pupils.

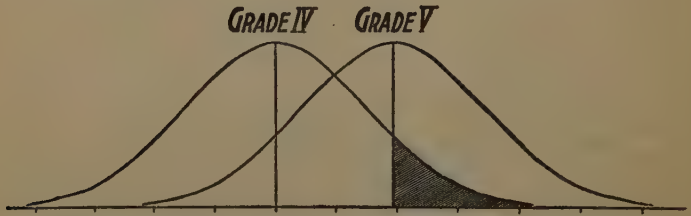


FIG. 48. — The overlapping of distribution of scores for the fourth grade on that for the fifth grade. The shaded portion of the fourth grade distribution (9.6%) represents higher scores than the median fifth-grade pupil attained.

Any measure of dispersion may be used in this manner, although the median deviation (P. E.) and the standard deviation (S. D.) have been used most frequently. The group may contain pupils of the same age, or it may be composed of the pupils in the same grade.

Where more than one such group of pupils is employed in deriving a scale, the value of each element in the test is first determined for each separate group, then the distance between groups is measured, and at last the final values for a given item are made up by referring each of these group values and differences to an arbitrarily chosen zero or reference point. As measured by results from fourth-grade pupils, for example, a problem may be 5.5 P. E.

above the chosen zero point, while by the results from fifth-grade pupils it may be 5.7 P. E. above zero. These values obtained from the separate distributions may be averaged in any sensible manner that is convenient. This procedure involves the assumption that each group is distributed according to the normal surface of frequency, and that the measure of dispersion in one group is equal to that in any other group similarly constituted. These assumptions are useful and are probably not far wrong, especially when large numbers of pupils are included in each grade or group.

If the separate groups are of this character, however, the total distribution obtained by combining all ages and all grades can hardly be represented by a normal surface of frequency. A long and relatively flat curve would probably represent such a grouping of scores for all ages and grades. The shape of such a curve would depend to a large extent on the proportion of pupils from each grade. It is much more in accord with known facts to make the assumptions regarding separate grade groups and to follow the procedure outlined in the above paragraph.

A question has recently been raised concerning the advantage of establishing the difficulty value in each school grade of each problem to be used in a scale.[1] It has been suggested that all scales for difficulty be standardized on the results obtained from twelve-year-old children, and that the exact amount of difference between the difficulties of various questions need not be determined. The answer to this suggestion is to be found in the degree of scientific accuracy desired, and in the ability level for

[1] W. A. McCall, *How to Measure in Education*, pp. 272–306.

which the scale is being prepared. If only pupils from ten to fourteen years of age are to be measured by the scale, the twelve-year-old group may be sufficient for purposes of determining the difficulties of the problems to be used. If pupils in the lower grades are to be measured, however, younger pupils should also be used in determining the values of the test elements. The relative difficulty of a question is not always the same with eight-year-old pupils as with twelve-year-olds. In any case, it is certainly worth the time and effort required to determine as accurately as possible the degree of difficulty in each element of any scale that is to be used extensively in evaluating the abilities of boys and girls.

Converting percentages of judgments of relative merit into P. E. locations. — The discussion thus far in this chapter has dealt with scales for measuring the difficulty of questions answered or of problems solved. The assumption has been that the problem most frequently solved is the easiest problem, and that *equally often solved problems are equal in difficulty*, unless always solved or never solved. Each performance of the child in such tests is to be scored either as "right" or as "wrong." There are certain situations, however, where the child's work cannot be scored as "success" or "failure," but must be rated according to its quality. In such cases the judgment of the teacher or judge needs to be standardized and rendered objective by the use of a scale for quality. The handwriting scales, composition scales, drawing scales, and scales in other branches of art are of this type, — measuring the amount of quality rather than its mere presence or absence.

The essential theory on which scales for relative merit

are built is that *equally often noticed differences in merit are equal*, unless always noticed or never noticed. The proof that one sample of an artistic product is better than another sample is that a majority of those who are competent to judge quality in that field declare it to be better. If one assumes that competent judges would vary just as much on one sample as on another, and that the form of the distribution would in each case follow the normal curve of frequency, Table LXIII may be used for converting the percentages of "better" judgments into P. E. units of difference.

In the preparation of his scale for the quality of English compositions, for example, Dr. Hillegas [1] had each of two hundred and two judges to arrange a set of English compositions in the order of their merit. One of these compositions, entitled "Going Down with Victory," may be designated *a*, and another, which was on the topic, "Ichabod Crane," may be called *b*. A hundred and fifty of the two hundred and two judges rated *a* as better than *b*, while fifty-two rated *b* as better than *a*. The proportion of the *b* distribution of judgments which was below *a* was therefore 74.3% ($\frac{150}{202} = .743$). Table LXIII shows that such a percentage indicates a difference of .97 P. E. in favor of *a*. Since the value of *b* had already been established as 6.75 P. E. above the zero composition, the *a* composition was given a value just .97 P. E. higher ($6.75 + .97 = 7.72$). The relative merits of the other samples were found in exactly the same manner.

An additional illustration may be taken from the field

[1] M. B. Hillegas, *A Scale for the Measurement of Quality in English Composition by Young People*, Teachers College, Columbia University.

of music appreciation. The writer desired to discover which of two phonograph records of instrumental music was better, and how much better the one was than the other. Both records were played in the presence of ninety intelligent adults who had more than the average amount of interest in good music. For one-half of the judges, record *b* was played before record *a*, while for the other half, *a* was played before *b*. Sixty-four of the ninety thought *a* better than *b*, while twenty-six thought *b* better than

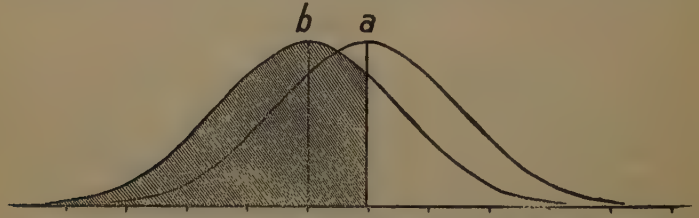


FIG. 49. — Overlapping of distributions of judgments on two English compositions. The shaded area represents 74.3% of the *b* distribution as being below the median of the *a* distribution.

a. In other words, 71.1% of the distribution for *a* was above the median of the distribution for *b*. Table XLIII shows that 71.1% is equivalent to a difference of .82 P. E. between the medians of the two distributions. Phonograph record *a* was therefore recorded as being .82 P. E. better than record *b*, in so far as the P. E. of this group of judges can be taken as a valid measure of the dispersion of judgments of competent persons in general.

The same general procedure may be used in determining the amount of difference in quality between any two samples of the same type of product. The percentage of the judges rating one sample as better than the other may be turned into P. E. distances between the medians of the

two distributions, by means of Table LXIII, or into S. D. differences by means of corresponding tables provided in statistical handbooks. The assumptions on which the procedure is based seem in the main to be sound, although there is evidence to show that the judges are more variable on some samples than on others.[1] In order that verification of such a scale may be made by other investigators at a later date, it is necessary that a most careful description be given of the manner in which the group of judges was selected; otherwise those who try to obtain a comparable group may find that their judges have a greater or a smaller dispersion than the group used originally in making up the scale. The writer has found that a few hours of intensive training will reduce the dispersion of a group of English composition judges to such an extent that their Median Deviation is only about five-eighths its original size.

Summary. — There is need for some permanent means of evaluating the relative difficulties of problems, for unequal intervals between elements in an educational scale are subject to even greater misinterpretations than are uneven steps in more common scales. The amount of dispersion in a group of pupils having the same age, or having the same school grade, is the most constant thing about such a group. Differences in difficulty have therefore been measured in terms of the P. E. or the S. D. of grade groups. The procedure involves certain general assumptions, but these seem to be entirely practical, even though there may occasionally be slight errors in their application to specific situations. Differences in the

[1] See pp. 69 ff., in M. R. Trabue, "Supplementing the Hillegas Scale," *Teachers College Record*, January, 1917.

quality of the artistic products of pupils are measured by objective scales, which have been constructed on the assumption that judgments of quality are distributed about the actual quality of the sample in a manner that can best be represented by the normal surface of frequency. The P. E. of a distribution of such judgments is the most common unit for measuring the amount of such quality.

QUESTIONS AND EXERCISES

1. If it were necessary to choose between a test which gave a false impression of the amount of achievement and an alternative test which gave a correct measure of the abilities of the pupils, but on which there were relatively few comparative scores, which would you choose? Why?

2. Suppose one had a series of posts, the first of which was 5 ft. high, the next 5 ft. 1 in., the next 5 ft. 2 in., and so on. Why would it be incorrect, after comparing a man's height with this series of posts, beginning with the shortest and taking them in order, to assign him as a final score the sum of the heights of all the posts which were shorter than he?

3. How many P. E. units more difficult is a problem solved by 5% than a problem solved by 50% of a class? How much more difficult is a problem solved by 35% of a class than another problem solved by 85%?

4. What is the P. E. difference between the medians of two groups of pupils, if only 20% of one group make as high a score as the median pupil in the other group? What assumptions are involved in your reply?

5. In a group of 200 average persons viewing two different moving-picture exhibitions, 165 thought the first exhibition was better than the second, while 35 thought the second exhibition was better than the first. How much better was the first than the second?

6. Assuming that ability to distinguish differences in virtue is distributed according to the normal surface of frequency, how much

worse is a crime that is judged worse than another by 95% of the intelligent people in a community?

REFERENCES FOR READING

L. P. Ayres, *A Measuring Scale for Ability in Spelling*, Russell Sage Foundation.

B. R. Buckingham, *Spelling Ability: Its Measurements and Distribution*, Teachers College, Columbia University, Contributions to Education, No. 59.

M. B. Hillegas, "Scale for the Measurement of Quality in English Compositions by Young People," *Teachers College Record*, Vol. 13, No. 4.

W. A. McCall, *How to Measure in Education*, Chapters IX and X.

E. L. Thorndike, "Handwriting," *Teachers College Record*, Vol. II, No. 2.

M. R. Trabue, *Completion Tests Language Scales*, Teachers College, Columbia University, Contributions to Education, No. 77.

CHAPTER XXI

PROSPECTS IN EDUCATIONAL MEASUREMENTS

SUMMARIZATION of a few of the outstanding effects of the measurement of results in education is entirely possible, although any one writer's account of the situation is certain to be fragmentary. The variety and number of problems for which measurements of mental traits aid in finding solutions is so great that no single individual is likely to be interested in all of them. The discussion which follows, therefore, concerns only a few of the tendencies and prospects which seem worthy of more attention than they have thus far received.

The question of the permanence of educational measurements need not be raised. The principles involved are as vital and as lasting in the field of education as they are in such fields as medicine, economics, or biology. The future will undoubtedly see radical changes and improvements in the measuring instruments, in the knowledge and skill of those who apply them, and in the effectiveness of the educational treatment that follows the testing. But the discontinuance of objective measurements of results, in a field in which change and improvement are the chief objectives for all activity, is impossible.

Probable changes in the measuring instruments. — It is probable that the measuring instruments to be developed in the future will tend to be more highly specialized

than those thus far produced. In the place of attempting to measure the *general* result obtained in a given subject, an increasing number of attempts will probably be made to measure various *specific* results. This will be especially true in the measurements intended primarily to assist teachers in the analysis of their instructional problems. It is usually sufficient for the purposes of administrative officers to know whether the general result in a given subject is up to standard. But the teacher and supervisor want to know something about specific results in more limited fields. For their purposes it is important to know, for example, both how rapidly and how well pupils write. The intelligent teacher is not even satisfied to learn that a pupil writes less well than he should. She wants also to know in what respects the child's writing is poor, in order that she may help him in exactly those places where his need is greatest.

As interest in the measurement of results passes from the administrative officers to the teachers, there will be an increasing demand for specific tests to reveal facts regarding various definite phases of the abilities of each pupil. This does not mean that the present scales will become useless. There will still be need for checking up the general condition of a pupil with respect to important abilities, and for relating these general findings to the results obtained from specific tests.

Teachers will employ objective measures of various conditions regularly, just as physicians have come to use the clinical thermometer and the pulse-count as means for measuring the symptom called "fever." Relatively few physicians would make a final diagnosis and pronounce

a person ill of "appendicitis," however, if there was no other evidence than that the patient had a temperature of 101 degrees. The exact location of any pains, the condition of the digestive system, and a large number of other symptoms are all considered *in relation to each other* before a medical diagnosis is reached by a careful physician. Successful teachers will be just as thorough in their study of the scores of a pupil. The measurements themselves will never improve the pupil's situation. But careful study of the relationships of the results obtained from tests will be of immense value to the specialist in educational practice, for these relationships will provide the basis for discovering the exact nature of the pupil's educational needs.

Improving educational measurers. — Teachers will undoubtedly make more and more use of exactly measured facts regarding their pupils, although it is by no means certain that they will themselves become specialists in the administration of all these tests. Not all physicians are expert in the operation of an X-ray machine, and many of them do not attempt to make microscopic or chemical analyses of body tissues and secretions, and yet all of them use the findings of such scientific tests. It is probable that, in a similar manner, all good teachers will learn to employ the measured results of pupil achievements, and perhaps to administer some of the more simple tests. But the chances are that the highly specialized technique of measurements will be mastered only by those who plan to give their entire time and attention to such work.

It will be the duty of specialists in measurements to supervise teachers in their uses of measured facts as well

as to supply these facts. If this is not done, many teachers will fail to recognize the limited reliability of certain test scores and will place entirely too much dependence on a single index of capacity. The specialist in measurements must therefore know educational work from a practical as well as from a theoretical standpoint. He must be able to see the educational limitations of his results as well as their statistical foundations. It must be possible for him to distinguish clearly the type of test that will most appropriately measure the thing the teacher is trying to give her pupils. There have been entirely too many attempts to measure power in a given subject by means of rate tests, and to measure rate by means of tests that involve many different degrees of difficulty.

The specialist in measurements will be called upon to plan and to direct experiments to determine the relative efficiency of various methods of instruction, alternative schemes of grouping pupils for study, different textbooks, and the like. He must therefore be characterized by an impartial, open-minded, scientific attitude. He must know how to safeguard his inquiries from such insidious fallacies as that of "the selected group." He must have native skill and good training in the presentation of his results, in order that they may be easily understood by teachers and laymen. Psychological training is quite as important a requirement for the specialist in educational measurements as is training in the field of education itself.

Effects of educational measurements on courses of study. — The effect of measurements upon courses of study depends to a large extent upon the thoughtfulness and insight of the teachers who are administering the

courses. When a test is given to the pupils of a teacher who does not think deeply and who does not have good supervision, she frequently concludes that the type of materials used in the test are the type she should be employing in her teaching. In perhaps twenty per cent of the classrooms in which the writer a few years ago gave a language completion test, the teachers began immediately thereafter to instruct their pupils in the art of writing appropriate words on blank spaces in sentences. In a still larger percentage of the rooms in which Courtis Arithmetic Tests are given, the teachers undertake at once to increase the speed at which their pupils do the fundamental operations in arithmetic.

This tendency to emphasize the process which has been measured is entirely natural. In a large proportion of her work the average teacher has had the feeling that she was dealing with intangible things and that there was no way of knowing when she had completed her task. It is only natural that when the means for doing so are made available, she should attempt to prove her ability to teach. Teachers have sought every imaginable means of demonstrating that they were doing good work. To drill pupils on the things for which they are later to be tested, especially when the tests are labeled "standard," is an opportunity no ambitious teacher is likely to overlook.

The effect of giving a certain type of test is sometimes to fix that type of exercise as a part of the regular routine of instruction. Giving a test in which the emphasis is on speed may tend to crystallize the course of study and to direct its chief attention to the development of speed. It requires real professional insight to decide when a test is

worth imitating and when it has value only as a test. Both teachers and directors of measurements should beware of this tendency for the teaching program to be determined in part by the testing program, and should safeguard the course of study by choosing their tests wisely and by warning thoughtless teachers of the danger.

When the more thoughtful teachers study the results of the objective tests given to their pupils, the effect upon the course of study is often to make it more rather than less elastic. If the relations between a child's intelligence test score and his score in algebra, for example, indicate that he is doing his utmost to succeed but is nevertheless failing to master the subject, the intelligent teacher may and often does begin to seek other subjects in which the pupil's efforts might be rewarded with "Success."

Probably no other agency has been more powerful than the measurement of results in calling attention to the injustice of the traditional curriculum. The state provides at public expense a course leading to the "learned professions." All pupils attempt this course, for the reason that there is no other provided, but relatively few remain in school from year to year (beyond the compulsory age-limit) except those whose abilities lead them to be interested in the abstract type of work provided in the school. Instead of training each child to do the service to the community which his mental inheritance makes it possible for him to render effectively, the schools have too frequently selected and retained for education at public expense only those few pupils whose interests and abilities enabled them to advance along academic lines. The objective methods of measuring results have been partly responsible for

bringing these facts to light, and thus indirectly for bringing about the tendency to broaden the curriculum and to provide training for all types of abilities.

Effects of educational measurements on teaching methods and supervision. — Teachers have been told by one publisher that the A——Method was most effective and by another that the B —— Method was the best. One supervisor supported the claims of the A —— Method, while another supported the B —— Method. It is not surprising that many teachers came to the conclusion that there was little real difference in the effectiveness of the different methods. The method adopted has in many cases been dependent upon the relative abilities of the salesmen sent out by the different publishing houses.

Educational measurements of the modern objective type have at last made it possible to determine with some degree of confidence the relative effectiveness of various methods of accomplishing a desired educational result. It is true that no single test is sufficiently broad in its scope to measure all the results sought for in any branch of school work. But the information and power measured by the test constitute a considerable part of the result sought for by the average teacher. As more and more experience is gained with measurements of those phases of ability that are now measurable, new elements in the objectives of instruction will be distinguished and subjected to measurement. Teaching and supervision will gradually be based to a larger and larger extent upon well-demonstrated facts, with less and less dependence upon unproved beliefs and prejudices. Teaching and supervision will become a science as well as an art.

The writer believes that the more exact statements of the nature of the teacher's task will make it possible to be more certain of success or of failure in its accomplishment. As soon as it becomes possible to demonstrate objectively to boards of education that one teacher is more highly successful than another, these boards should be willing to pay more adequate salaries for efficient teachers, and men and women of increasingly high qualifications may be attracted to the profession. A keener professional conscience and a more permanent group of teachers should be one of the ultimate results of a wider and more discriminating use of educational measurements.

Effects of measurements on school examinations. — The usual procedure of the teacher who wishes to test the effectiveness of her instruction is to set an examination of five or ten questions, in which she asks the pupils to "discuss," "explain," "tell," or "describe" some of the situations or events that have been studied. The resulting essays of the pupils are read by the teacher and given percentage marks or letter grades indicating her estimates of their quality. One teacher may rate examination papers according to the number of facts the pupils remember and mention, another may rate them by the way in which the pupils have organized the facts remembered, and still another may rate the essays by the degree to which they reveal an appreciation of the significance of the facts related. In practically all cases, however, a well written paper is given higher credit than one of like quality which is expressed in poor English or in poor handwriting.

School examinations have recently tended to be modified through the adoption of certain parts of the technique

used with standard educational tests. One of the great changes has been the tendency to substitute several specific short-answer questions for the general discussion type of question. The chief advantages of the larger number of short, specific questions over the smaller number of long, general questions are to be found in the greater objectivity of the short-answer questions. Instead of trying to rate the quality of a long discussion written by the pupil, the teacher needs only to check the accuracy of specific words or check marks. Two teachers rating the same essay examination might differ widely in their estimates of its worth, although they would agree closely in their ratings of a paper containing a hundred items of fact or good judgment. Having no opportunity to hide his ignorance in a long essay of well-written sentences and pointless paragraphs, the bright pupil who has not mastered the subject finds it necessary to allow his answers to stand "out in the open," where they may be evaluated on the same basis as the answers given by less capable pupils who have worked faithfully.

As examples of the short-answer type of examination one may take the following questions concerning the trial scene in the *Merchant of Venice*.

1. Underline the word which most nearly describes Shylock's original attitude toward the court.

REVERENT CONCILIATORY SORROWFUL THREATENING OBEDIENT

2. Underline the word which most accurately describes the merchant's feeling regarding the probable outcome of the trial.

HOPEFUL FURIOUS ANXIOUS RESIGNED REBELLIOUS

3. Underscore the excuse sent by the wise Bellario for not being able to attend the trial.

HE WAS DRUNK ILL ARRESTED INJURED TIRED

4. What was the real name of the young judge who asked Shylock to have mercy on the merchant? . . . . . .
5. By the name of what Old Testament character did Shylock call the young judge? . . . . . .
6. Antonio's bond awarded to Shylock a . . . . . of . . . . . (Supply the missing words in the above sentence.)

Put a T before each statement that is true, and an F before each statement that is false.

7. Shylock asked for and was offered justice.
8. This story shows that all Jews are hateful.
9. Shylock gave his property to his daughter and her husband.

Questions 1, 2, and 3 illustrate a type with which it is possible to test the pupil's ability to think clearly about character and motives as well as to test his memory for facts. With questions of the type illustrated in 7, 8, and 9 it is possible to cover an enormous number of details in a very short examination period, although the teacher cannot be certain that any child actually knows the correct answer to any particular question, for only two alternatives are allowed. In a " true-false " or " yes-no " examination, the pupil who answers half the questions correctly probably knows nothing about the subject, for half the replies would be correct by mere chance, even if one guessed at all of them. Questions of the type illustrated by 4, 5, and 6 are excellent when it is desired to test the child's memory for exact facts and events.

The chief objection to the short-answer questions is that, unless the teacher uses great care in the preparation of her questions, too large a proportion of them will involve only memory for details, and too few will require any thorough organization or interpretation of the facts.

The answer to this objection is that it is possible to prepare questions, especially of the type illustrated by 1, 2, and 3, which will *require* large units of thought and proper subordination of insignificant details. The usual essay type of examination allows the pupil to *choose* whether his answer will be superficial or profound. "Describe the trial scene," or "How did Shylock act at the trial?" may be answered on almost any level the pupil wishes to choose. Although it requires much more time and careful thought on the part of the teacher to prepare fifty good short-answer questions than to prepare ten of the usual examination questions, the marking of the short-answer papers may be accomplished in less than one-fourth the time required to judge the quality of an essay-type examination taking the same amount of the pupil's time. The short-answer test shifts the pupil's efforts from the mechanics of writing to concentrated thinking about the subject, and it shifts the teacher's time from the reading of examination papers to the preparation of test questions.

The larger number of questions possible in a given length of time enables the teacher to test a given field more thoroughly with the short-answer examination than with the traditional examination. Not only may a larger proportion of the field be covered in a given length of time, but specific information or ability may be distinguished more clearly from general abilities. Such irrelevant matters as legibility of handwriting, clearness of English, force and enthusiasm of expression, and personality of the pupil have less chance to modify the teacher's judgment of the results when the short-answer type of test has been used. Although the expense of multigraphing or printing the test questions may

be higher than in the case of the usual examination, there is real economy in giving a routine mechanical task to a clerk, if by doing so teachers are relieved of the monotony of reading long essays which give them less adequate information about a much smaller portion of the field.

One of the most important advantages of examinations made up of questions such as the first six illustrated above is that the teacher may more readily check up her own success in teaching specific habits, facts, or appreciations. By tabulating the percentage of right responses to each question asked, the teacher may discover which items have been most generally learned. By repeating year after year a few of the very same questions and tabulating the pupils' responses each year on each question, she may learn that she has greater success in teaching by one method than by another. The possibility of such a diagnosis of her instructional problems by the teacher herself, or by her supervisor, is one of the most important reasons for having written examinations of the pupils.

Most pupils would rather answer fifty questions specifically than ten questions by long discussions. The excuse "I knew it but could not express it" will not often be made regarding the short-answer type of examination. Pupils not only enjoy taking such a test, but they are more easily convinced of their short-comings by it. Even children have greater confidence in the abilities of teachers to judge the rightness or the wrongness of specific answers to pointed questions than they have in the abilities of these same teachers to judge how many points out of ten should be allowed for an imperfect answer to a general question.

The effects of standard tests on scholarship marks. — Standard tests have supplied the exact objective measurements of ability upon which the grouping together of those of like abilities must depend. They have made it possible to state the objectives of a course of study in such a form that the teacher can know when a pupil has achieved them. Measurements have demonstrated the impossibility of putting all pupils through the course of study designed for the craftsmen who deal in words and ideas. They have pointed out the folly of attempting to make a single scholarship grade tell at one and the same time the pupil's fitness for promotion, the amount of his achievement, his standing as compared to other pupils in the class, his earnestness of effort, and his ability at one time as compared with his ability at some previous time.

As was pointed out in Chapter II, the most fundamental problem with regard to scholarship marks concerns the purpose for which they are to be used. If a pupil's mark in arithmetic on his report card is to indicate to the parent the relative position of the pupil in his arithmetic class, it is not wise to use scores from standard arithmetic tests as marks. Parents cannot be expected to become experts in educational measurements. On the other hand, unless the records kept in the school furnish something more objective than a statement of the relative standing of each pupil in a group, these records will not be of service in determining the effectiveness of different methods of teaching, nor will they offer the maximum of help in guiding the pupils educationally and vocationally. The situation therefore resolves itself into the necessity for at least two types of scholarship marks, the one to stand as an objective

record of the pupil's work in the school, and the other to serve each month as an indication to parents of the relative standing of the pupil in his class.

The best objective records for the use of the school authorities are the scores that may be obtained from the standard tests and from the short-answer examinations which have been discussed above. The marks for use on reports to parents may be obtained indirectly in various manners from the same sources and from the pupil's daily work, the basis depending on the characteristic which it seems best to let these marks indicate. There is much to be said in favor of letting the mark sent home on the report card indicate the relative amount of effort the pupil has put on his work. A child who was unusually capable in reading might put forth relatively less effort and deserve a lower mark under this plan than a very slow pupil who had more nearly done his best. Pupils can never be made equal in their abilities. Both teacher and parent should, however, inspire each pupil to attain as nearly as possible his maximum development. Where the school decides to reward effort rather than actual achievement, scholarship grades for the report cards might well be obtained from the various accomplishment ratios earned by the pupil.

Whether the scholarship marks sent to parents depend on accomplishment ratios or on absolute amounts of ability, there is little reason to allow them to influence a pupil's classification. Pupils should always be grouped in such a manner as will permit those of like interests and abilities to work together, regardless of the scholarship marks that have been reported to their parents. Failure of a child to be promoted indicates either that the school has not ad-

justed its courses to the ability of the child or that the teacher has failed to stimulate in the pupil the effort of which he is capable. Readjustments in the grouping of pupils should be made whenever necessary.

When one has selected the basis upon which scholarship marks are to be given, the one important question remaining is that concerning the proportion of marks of each size. Those who have studied the matter carefully have come to the conclusion that there is no advantage in trying to distinguish as many as 25 grades of ability between a passing mark (75%) and perfection (100%). Five different grades are about as many as are necessary for the purposes of reports to parents, and it makes little difference whether these be called " A, B, C, D, and E," or " 1, 2, 3, 4, and 5." The important question for the parents is " How does my child compare with the others this month? " In order to answer this question intelligently, the parent must know what proportion of the class is given each grade.

The answer to this question of the relative frequency of each grade will to a large extent be an arbitrary decision. Some writers advocate giving the best 5% of the pupils a mark of A, and the lowest 5% a mark of E. Others think that there should be a larger proportion of A and E marks. There are some who would give the middle 50% of the pupils a mark of C, while there are others who would give C to the middle 30% only. There is substantial agreement that one should give approximately the same proportion of A marks as E marks, and the same proportion of B's as of D's. In other words, scholarship marks should be distributed symmetrically about their average. The normal curve of frequency probably applies quite well in most

classes to the distributions of both abilities and accomplishment ratios.

As soon as one assumes that the marks he is to assign should conform in their distribution to the normal surface of frequency, he may find help in Table LXIII. If one chooses to give the middle 50% of the pupils a C grade, this would mean that all pupils between the 25 percentile and the 75 percentile would be marked C. The P. E. difference between the 25 percentile and the 75 percentile is 2.0 units, according to Table LXIII. If the difference between the best and the poorest C grade is to be 2.0 P. E., then there should be just as much difference between the best B and the poorest B, and just as much between the best D and the poorest D. According to Table LXIII, in order to get 2.0 P. E. above the 75 percentile one must go to the 98 percentile, and similarly to get 2.0 P. E. below the 25 percentile one must go down to the 2 percentile. If the middle half of the class is to be given a grade of C, therefore, it would be fair to make the following distribution of grades:

A	Best 2%
B	Next 23%
C	Middle 50%
D	Next 23%
E	Lowest 2%

In case one does not wish such a small percentage of the extreme grades, he can prepare a different scheme. Suppose, for example, that one wishes to assign the lowest 10% a grade of E and the highest 10% a grade of A. The difference between the 10 percentile and the 90 percentile, according to Table LXIII, is 3.8 P. E. If this difference is divided into three equal parts of approximately 1.27

P. E. each, the two dividing points would be at approximately the 34 percentile and the 66 percentile. The distribution would then be as follows:

A	Best 10%
B	Next 24%
C	Middle 32%
D	Next 24%
E	Lowest 10%

Figure 50 illustrates the proportion of the group to be given each mark under the two plans tabulated above. Other equally good distributions of proportions may be worked out in a similar manner by the use of Table LXIII.

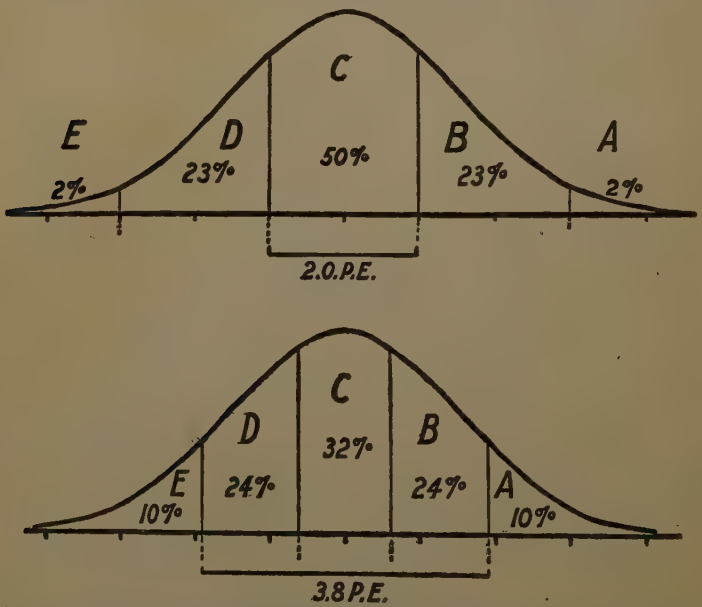


FIG. 50. — Illustrative distributions of scholarship marks. Given the percentage of marks that is to be of any particular grade, the percentages for each other grade follow from the nature of the normal surface of frequency.

Grades may then be assigned in their proper proportions, whether the percentile points used as a basis are scores on some standard educational test, ratings in order of merit for some important characteristic, scores in a short-answer examination, or accomplishment ratios worked out from educational and intelligence tests. In any case, the parent would have a reasonably correct understanding of the significance of a change from D to B in the monthly grade of his child.

Summary. — The fact that education is a matter of changes and improvements makes it certain that objective measurements will continue to be used in schools with ever-increasing accuracy and effectiveness. Specialists in educational practice will learn to interpret more confidently the relations found between the results obtained from various measurements, and to make on these bases more adequate adjustments of the instruction to the interests and abilities of the pupils. The measuring instruments themselves will no doubt become more and more accurate and will probably measure more limited portions of the fields to which they apply.

The general effect of measurements will be to make education more scientific and more democratic, through the simple process of stating more exactly the results that are being obtained under present methods and organization. By demonstrating the superior effectiveness of certain methods of accomplishing desired results, measurements will make possible a real science of teaching and supervision. By proving the impossibility of teaching some of the present school subjects effectively to certain groups of pupils, new activities will be suggested for addition to the

curriculum, in order that all types of citizens may obtain genuine profit from the schools maintained at public expense.

QUESTIONS AND EXERCISES

1. For what reasons are exact measurements certain to be continued in chemistry, economics, and education?

2. Why is it probable that measuring scales will in the future concern themselves more closely with the small features in a total result, rather than with the larger, general situation?

3. On what type of information does the trained physician make a clinical diagnosis? How would an expert teacher make a diagnosis of the educational ailments of a particular pupil?

4. What would you expect a diagnostic series of tests in arithmetic to contain? Would it be better to have a single final score in a diagnostic test, or several separate scores?

5. In what respects would training in educational psychology be of assistance to the expert in educational measurements?

6. Under what conditions might educational measurements have a detrimental effect on the curriculum of a school?

7. Why does the result of a test of 20 questions, in which one is simply to mark whether a given statement is true or false, fail to have a high correlation with the intellectual abilities of the pupils? Why is the true-false test less accurate — page for page — than tests in which the student writes one word as an answer to each question?

8. In order to distribute class marks symmetrically and equitably with C as an average, what percentage of all marks should be of each grade from A to E, if 8% of the class are to be given A's?

9. Is it desirable that the grades sent home to parents, on pupils' report cards, should have any real meaning?

REFERENCES FOR READING

C. A. Gregory, *Fundamentals of Educational Measurement*, Chapters II and V.

W. A. McCall, *How to Measure in Education*, Chapter I.

W. S. Monroe, *Theory of Educational Measurements*, Chapter II.

INDEX

Ability: evidence of, 19, 282; a variable quantity, 19 ff.
Accomplishment quotient *or* ratio, 445.
Adults' *vs.* children's scales, 322–324.
Age: calculation of average, 89 ff.; comparisons by, 169; educational, 445; effect of, on scores of pupils in same grade, 167–169; how to secure records of, 84 ff.; importance of accurate statements of, 80, 93; inaccuracies in statements of, 81 ff.; table for computing, 86.
Amount *vs.* quality of educational products, 327–331.
Arithmetic tests: Buckingham Scale for Problems in Arithmetic, 307–312, 323, 334; Cleveland Survey Arithmetic Tests, 287–292, 330; Courtis Standard Research Tests: Arithmetic, Series B, 76, 275–287, 301, 303, 326, 330; Stone Reasoning Tests, 72, 73, 301–307; Woody Arithmetic Scales, 294–301.
Army intelligence tests, 421–423.
Arrays of measures, 99, 202.
Average, see *Mean*, *Median*, and *Mode;* also page 151.
Average measures: classification of, 198; comparisons among, 224.
Ayres, Leonard, 76, 77, 137.
Ayres Handwriting Scales, 76, 77, 191 ff.
Ayres Spelling Scales, 76, 137, 140, 141.

Ballou, Frank W., 379.
Bimodal distributions, 205 ff.
Binet, Alfred, 413 ff.
Binet-Simon Tests, 413 ff.
Briggs English Form Tests, 359.
Buckingham, B. R., 75, 307, 334.
Buckingham Extension of the Ayres Spelling Scale, 133, 135, 136, 138.
Buckingham Scale for Problems in Arithmetic, 307–312, 323, 334.
Burgess, M. A., 26; quoted regarding time for power tests, 353 ff.
Burgess Scales for Silent Reading, 353 ff.

Cattell, J. McK., 413.
Central tendencies, see *Mean*, *Median*, and *Mode.*
Chadwick, E. B.; quoted regarding Fisher's "Scale-Book," 70, 71.
Chance, effects of, 236–241.
Children's *vs.* adults' scales, 322–324.
Classifications: of mental tests, 314 ff.; of pupils, 436–439, 441–444.
Cleveland Survey Arithmetic Tests, 287–292, 330.
Coefficient of correlation, see *Correlation coefficient.*
Comparisons, necessity of proper conditions for, 324–327, 331–333.
Comparisons of distributions by: amount of overlapping, 161–164; average scores, 97 ff.; decile points, 116, 118; frequency tables and surfaces, 100 ff.; middle-fifty-per-cent ranges, 115; percentile curves, 164 ff.; percentile values, 111–116; quartile points, 117; seriation, 99.
Comparisons of pupils of same age rather than of same grade, 169.

Composition scales, English: Briggs English Form Tests, 359; Harvard-Newton Composition Scales, 379, 380; Hillegas Composition Scale, 74, 364 ff., 451; Hudelson English Composition Scale, see *Virginia Supplement;* Hudelson Typical Composition Ability Scale, 371 ff.; Lewis Composition Scales, 381, 382; Minnesota English Composition Scales, 380, 381; Nassau County Supplement to Hillegas Scale, 368 ff.; Thorndike's Extension of the Hillegas Scale, 367 ff.; Virginia Supplement to Hillegas Scale, 371 ff.; Willing Scale, 381.
Conditions affecting results, 131 ff., 182–185, 325–327.
Continuity: in scales for age, 91 ff.; in scales for spelling, 145–147; of time and ability, 91–93.
Cornman, O. P., 71, 132.
Correlation coefficients: advantages of, 390–392; calculation of, 392–404; interpretation of, 406 ff.; a measure of reliability, 407–409; reliability of, 404–406.
Courses of study: narrowness of, 60 ff.; probable effects of educational measurements on, 471–474.
Courtis, S. A.: 76 ff.; 276 ff.; quoted regarding speed tests, 280, 281, 303, 357.
Courtis Standard Research Tests: Arithmetic, 76, 276 ff., 301, 303, 326, 330; Reading, No. 2, 357.
Cumulative frequency tables, see *Frequency table.*
Curriculum: narrowness of, 60 ff.; probable effects of measurements on, 471–474.

Dearborn, W. F., 424.
Decile points, 116–118.
Decile table, 166.
Derivation of scale values, 332 ff., 449 ff., 456–465.
Development of standard tests, 59 ff.
Difficulty: basis for measuring relative, 332 ff., 452–455, 461–465.
Dispersion: definition of, 253; measured by defined ranges, 115 ff., 155 ff., 254; measured by mean deviation, 260 ff.; measured by median deviation (P. E.), 257 ff.; measured by quartile deviation (Q), 254 ff.; measured by root-mean-square deviation (σ), 265 ff.; relations between various measures of, 267; as unit for measuring differences, 455 ff.; uses of measures of, 269–271.
Distribution: normal, see *Surface of frequency, normal;* of samples from total population, 241–244, 252 ff.; skewed, 244–246.
Dominant variable, 327, 332 ff.

Educational age, 445.
Educational measurements, early attempts, 62 ff., 69 ff.
Educational *vs.* intellectual scales, 318–320.
English composition scales, see *Composition scales, English.*
Examinations: as bases for school marks, 51 ff.; disagreements in marking, 53 ff.; probable effects of educational measurement on, 475–479.

Fisher, Rev. George, 70, 71.
Freeman (F. N.), Chart for Diagnosing Faults in Handwriting, 193 ff.
Frequency curve, definition, 235 ff.
Frequency polygon, 247.
Frequency, surface of, see *Surface of frequency.*

Frequency table: advantages of, 101 ff.; cumulative, 112 ff.; definition of, 100; methods of preparing, 106–111.

Galton, Francis, 413.
Gettysburg Edition of the Ayres Scale, 192, 193.
Goddard, Henry, 414.
Grades: scholarship, see *Scholarship marks;* assignment to, see *Classification of pupils.*
Grading, see *Rating;* also *Classification of pupils.*
Gray, W. S., Oral Reading Tests of, 345, 347 ff.
Gray Silent Reading Tests, 354–356.
Group *vs.* individual tests, 321 ff.

Haggerty, M. E., 424.
Haggerty Reading Examination, 358.
Handwriting: directions for rating quality of, 176 ff.; directions for standard test of, 173 ff.; influence of conditions on, 182 ff.; objectives for teachers of, 185–187; obtaining a sample of, 174–176; standard scores in, 181 ff.
Handwriting scales: Ayres Handwriting Scales, 76, 77, 191 ff., Gettysburg Edition, 192; Freeman Chart for Diagnosing Faults in Handwriting, 193 ff.; Starch Measuring Scale for Handwriting, 193, 391; Thorndike Scale, 73 ff., 175, 177 ff., 328.
Harvard-Newton Composition Scales, 379 ff.
Herring, John P., 419.
Herring's Revision of the Binet Scale, 419, 425.
Hillegas, Milo B., 74, 364, 451, 463.
Hillegas Composition Scale, 364 ff., 451.
Histogram, 103, 247.
History of development of standard tests, 59 ff.
Hollerith Sorting Machine, 109.
Hudelson, Earl, 371, 451.
Hudelson English Composition Scale, see *Virginia Supplement.*
Hudelson Typical Composition Ability Scale, 371.

Individual *vs.* group tests, 321 ff.
Intellectual *vs.* educational scales, 318–320.
Intelligence, nature of, 429–432.
Intelligence quotient, 416 ff.; as basis of classification, 441–443; as basis of vocational choice, 435.
Intelligence tests: characteristics measured by, 430–432; choice of, 425–427; in educational guidance, 432–436; in educational organization, 436–440.
Intelligence tests: Army, 421–423; Binet-Simon, 413 ff.; Dearborn, 424; Haggerty, 424; Herring's Revision of the Binet Scale, 419, 425; Mentimeters, 425; Miller, 424; National, 424; Otis, 420, 421; Pintner-Paterson Scale of Performance Tests, 418; Stanford Revision of the Binet Scale, 418 ff.; Terman Group Test of Mental Ability, 424; Yerkes-Bridges Point Scale, 417.
Interdependence of subjects, 320 ff.
I. Q., see *Intelligence quotient.*

Judd, Charles H., 288.

Kansas Silent Reading Tests, 358.
Kelley, T. L., 334.

Learning, evidences of, 18, 19.
Lewis, E. E., 382.

Lewis English Composition Scales, 381, 382.

Machines, Sorting, 109.
Marks, see *Scholarship marks*.
McCall, W. A., 351.
Mean, arithmetic: calculation of, with guessed mean, 225–227; calculation of, from tabulated data, 151–154, 190, 209–216; calculation of, from untabulated data, 206–209; definition, 151, 198; formula for calculation of, 222; a frequent error in calculation of, 89 ff.; inadequate as a measure of a group, 96–98.
Mean deviation: calculation of, 260–265; compared with other dispersion measures, 267; definition of, 260; uses of, 269–271.
Measurement: nature of, 26–28; amount of difficulty of, 332 ff., 456–461; amount of quality of, 462–465.
Measurements: advantages of, 28 ff.; early attempts, 63 ff., 69 ff.; qualifications of specialist in, 470 ff.; early standards of, 72.
Measuring instruments, probable changes in, 468–470.
Median: calculation of, from distributions, 123–125, 147–151, 189, 216–221, 384; definition of, 198; formula for calculation of, 223.
Median deviation: calculation of, 257–260; compared with other measures of dispersion, 267; definition of, 257; uses of, 269–271, 456–465.
Mental *vs.* physical scales, 318.
Mentimeters, 425.
Methods and supervision, probable effects of educational measurements on, 474 ff.
Middle-fifty-per-cent range: calculation of, 155–161, 191; as measure of dispersion, 115; as summary, 154 ff.
Miller, W. S., 424.
Minnesota English Composition Scales, 380 ff.
Mode: definition of, 198; determination of, 203 ff.
Monroe Standardized Silent Reading Test, 359.
Multimodal distribution, definition of, 205.

Nassau County Supplement to the Hillegas Scale, 368 ff.
National Intelligence Tests, 424.
Nature of Measurement, 26–28.

Objective *vs.* subjective scales, 316 ff.
Otis, Arthur S., 420.
Otis Tests, 420 ff.
Overlapping, as a means of comparison, 161 ff.

P. E., see *Median deviation*.
Pearson, Karl, 272.
Percentile curves, 111–114, 164 ff.
Percentile points: calculation, 120–126, 187–189, 382–385; comparisons by means of, 111–120.
Percentile values: definition of, 113; calculation of, 112 ff.
Physical *vs.* mental scales, 318.
Pintner-Paterson Scale of Performance Tests, 418.
Population, definition of, 241.
Power *vs.* rate, 282 ff., 332 ff.
Powers Sorting Machine, 109.
Pressey, L. W., 342.
Probable error, see *Median deviation*.
Processes *vs.* results, 324–327.

Q., see *Quartile deviation*.

Quality: basis for measuring relative, 452–455, 461–465; degrees of, 24; measuring differences in, 462 ff.
Quartile deviation: compared with other measures of dispersion, 267; as a measure of dispersion, 254–257; uses of, 269–271.
Quartile points: calculation of, 125, 384 ff.; definition of, 117.

Range desirable in scales, 322–324.
Rate tests: dangers of, 280–284; *vs.* power tests, 282, 336 ff.
Rating system, faults of, 49 ff.
Reading *vs.* other school subjects, 320 ff.
Reading ability: discussion of, 339 ff.; units of, 340–345.
Reading tests: Burgess Scales for Silent Reading, 353 ff.; Courtis Standard Research Tests, Silent Reading, No. 2, 357; Gray Oral Reading Test, 345 ff.; Gray Silent Reading Test, 354–356; Haggerty Reading Examination, 358; Kansas Silent Reading Tests, 358; Monroe Standardized Silent Reading Test, 359; Pressey and Skeel Reading Test, 342; Thorndike Test of Word Knowledge, 342; Thorndike Visual Vocabulary Scales, 342, 348–350; Thorndike-McCall Reading Scales, 350–353.
Relationship, measurements of, 389 ff.
Reliability of coefficient of correlation, 404–406.
Reliability of tests, 425 ff.; measured by correlation coefficient, 407–409.
Report cards, 35 ff., see also *Scholarship marks* and *Scholarship records*.
Reports to parents, purposes of, 36.
Results, conditions affecting, 131 ff., 182–185, 325–327.
Rice, J. M., 63 ff., 129.

Sampling, principle of, 241–244.
Scale and test, distinction between, 25 ff., 315 ff.
Scale values, derivation of, 332 ff., 449 ff., 456–465.
"Scale-book," Fisher's, 70 ff.
Scholarship marks: based on examinations, 51 ff.; indefinite significance of, 35 ff., influence of standard tests upon, 480–485; interpretation of, 41 ff.; lack of uniform standards in assigning, 43–49, 53–55; proper basis for, 482–485; purposes of, 36–38; scientific assignment of, 483–485.
Scholarship records, purposes of, 38–41.
School marks, see *Scholarship marks*.
Score-card, use of, 329 ff., 363.
Scores, assignment of, 333–336, see also *Handwriting scales*, *Reading tests*, etc.
Scores, standard, see *Arithmetic tests*, *Spelling tests*, etc.
Semi-interquartile range, see *Quartile deviation*.
Seriation, advantages of, 98–100.
Sigma, 265, see also *Standard deviation*.
Simon, T., 413.
Skeel, H. V., 342.
Skewness: measures of, 271–273; positive and negative, 244–246.
Sorting machines, description of use of, 109–111.
Specialist in measurements, qualifications of, 470 ff.
Spelling test: directions for administering and scoring a standard, 133 ff.; standard lists, advantages of, 140 ff.

Spelling tests: Ayres Spelling Scale, 76, 137, 140, 141; Buckingham Extension of the Ayres Spelling Scale, 133, 135, 136, 138; Iowa Spelling Scale, 143; Rice's spelling test, 64, 129; Springfield Tests, spelling list in, 67, 130; Standard Spelling Test Lists, 135, 136, directions for, 133 ff.
Springfield Tests, 66 ff., 130 ff.
Standard: a relative term, 25; scale, 25, 314; scales, the first, 73 ff.; tests, development of, 59 ff.; yardstick, 22 ff.
Standard deviation: calculation of, 266 ff.; compared with other measures of dispersion, 267; formula for, 267; sigma, 265; uses of, 269–271.
Stanford Revision of the Binet Scale, 418 ff., 425.
Stone, C. W., 73, 76, 303, 307.
Stone Reasoning Tests in Arithmetic, 72, 73, 301–307.
Strayer, George D., 77, 300.
Strayer-Engelhardt Score-Card for City School Buildings, 330.
Subject matter for school work: narrowness of, 60 ff.; probable effects of measurements on, 471–474.
Subjective *vs.* objective scales, 316 ff.
Summary of distribution: middle fifty per cent, 115, 154–161, 191; see also *Mean*, *Median*, *Mode*, *Percentile curve*, *Surface of frequency*.
Supervision, probable effects of educational measurement on, 474 ff.
Surface of frequency: normal, 232–235; representing distributions of scores, 103–106, 246–250; see also *Histogram*, and *Frequency polygon*.
Tabulating machines, 109–111.
Teachers' marks, see *Scholarship marks*.
Terman, Lewis M., 418, 420, 422, 424.
Test and scale, distinction between, 25 ff., 315 ff.
Tests, conditions of, affecting results, 131 ff., 182–185, 325–327.
Thorndike, E. L., 31, 72–74, 129, 334, 342, 350, 354, 424.
Thorndike Extension of the Hillegas Scale, 367 ff.
Thorndike Handwriting Scale, 73, 74, 175, 177–179, 181 ff., 188, 191 ff., 328.
Thorndike Test of Word Knowledge, 342.
Thorndike Visual Vocabulary Scales, 342, 348–350.
Thorndike-McCall Reading Scales, 350–353.
Time *vs.* difficulty of producing a result, 331–333.

Units, desirability of equal, 450 ff.

Van Wagenen, M. J., 380.
Variability, see *Dispersion*.
Variable, principle of the dominant, 324–327, 331, 333 ff.
Virginia Supplement to the Hillegas Scale, 371 ff.

Willing Scale, 381.
Woody, Clifford, 297 ff., 334.
Woody Arithmetic Scales, 294–301, 311.

Yerkes-Bridges Point Scale, 417.

Zero point, 298–300.

HETERICK MEMORIAL LIBRARY
371.26 T75 onuu
Trabue, Marion Rex/Measuring results in
3 5111 00057 2143

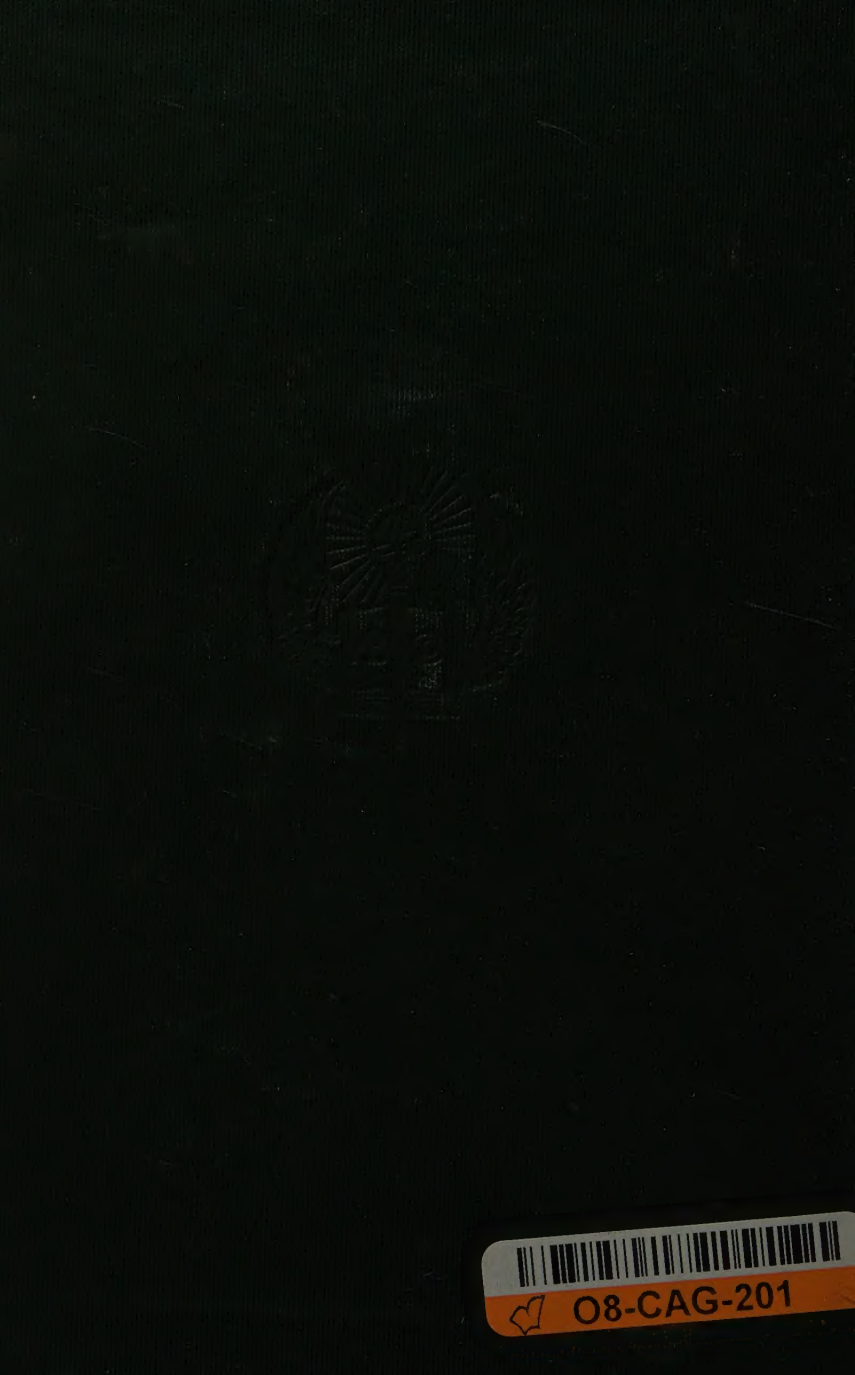
DATE DUE

FE 16 '81
MAY 15 2003
APR 25 2003
WITHDRAWN FROM
OHIO NORTHERN UNIVERSITY
LIBRARY

GAYLORD
PRINTED IN U.S.A.

25107

371.26
T75


08-CAG-201